27

Counts of Counts of Instances

Introduction 27-1
- basic counts can give structure 27-1
- species vs. instances 27-2
- mastering our techniques 27-3

27A. Raw counts sometimes work 27-4

27B. Root counts are usually best 27-10
- the comparison curve 27-18
- simplifying the arithmetic 27-18

27C. Further examples 27-24
- comparison with other fits 27-30

27D. Log counts are often needed 27-30
- sentence lengths 27-37a

27E. Reciprocal counts can be needed 27-43

27F. What next? 27-47

****

27H. How far have we come? 27-50

****

27K. Sometimes other expressions are needed (background) 27-51

27L. Poisson distributions 27-56

****

27P. Problems for chapter 27 27-63

Exh.	Page
A	
1	27-5
2	27-6
3	27-7/8
4	27-9a
B	
5	27-11
6	27-13
7	27-14
8	27-15
9	27-16
10	27-17
11	27-19/
12	27-21
13	27-22/
C	
14	27-25
15	27-26/
16	27-28
17	27-29
D	
18	27-31
19	27-32
20	27-33
21	27-34
22	27-35
23	27-36
24	27-38/
25	27-40
26	27-41
E	
27	27-44/
28	27-46
F	
H	
K	
29	27-52
30	27-53
31	27-54
32	27-55
L	
33	27-58/
34	27-60
35	27-61
P	
36	27-64
37	27-65
38	27-66
39	27-68
40	27-69
41	27-70
42	27-72

27

Counts of Counts of Instances

We are now ready to turn to the case where we are analyzing counts of counts, something that happens rather frequently.

* basic counts can give structure *

When we wish to analyze counts, the things that are counted all too often have no helpful structure -- no natural order, no cross-classification by two or more factors, no association with another ordered or numerical variable. This leaves us only two ways to make an analysis:

- to use the names of the things, which we can do effectively if both there are not too many things and their names are meaningful enough.
- to use the sizes of the counts themselves, clubbing together those things associated with the same basic count, and telling about them by saying how many there are.

To fall back on counts of basic counts might seem a counsel of desperation. It is not. As we shall see in many examples, it is often an effective way to get hold of significant information, though we cannot expect it to tell the whole story.

Once we have looked at -- and analyzed -- counts of basic counts, it is usually time to ask what could be done next -- what kind of information do we have about this sort of data that could be used -- perhaps in another study -- to go further.

* species vs. instances *

It will help us greatly to be able to distinguish two kinds of basic counts -- instances and species.

A man counting plants does one thing when he divides a field or forest into areas -- each of which is an instance -- and counts the number of plants -- perhaps all plants, perhaps all plants of a family, perhaps all plants of a species -- in each of these areas. If he does more field work of the same sort, he will study more instances and combine results by setting down new basic counts beside the old. The old instances will continue to appear with the old counts. He will have no extra difficulty with instances for which the count is zero. To each INSTANCE, one basic count.

A man counting plants in a field or forest does quite another thing when he counts the number of plants of each species -- or family, or size -- in a single suitable area. If he does more field work of the same sort, he will study more area and combine results by adding his new to the old. The old species will appear with new basic counts.

He may have a hard time with basic zeros. How many species did not occur? (Which species might have occured? How many species are there, anyway?) To each SPECIES, a basic count depending on how broad a look has been taken.

A man counting words per sentence is dealing with instances. A man counting occurrences of words in running text is dealing with species.

We shall deal with counts of species in chapter 27x, and with counts of instances in this one.

Instances are often like the "faceless mob", species are recognizable individuals.

Some patterns of basic counts are called J-shaped, because the smallest basic counts -- zero or one -- occur more frequently than any others, with frequency of occurrence declining as we look at larger and larger basic counts. Basic counts for species -- usually starting with ones, not zeros -- are often very J-shaped, with a very long tail. As a result they deserve -- and will get in chapter 27x -- special techniques of analysis.

* mastering our techniques *

When we deal with counts of basic counts, we have to deal with the idiosyncrasies of counts at two levels. At the upper level, our problems are not new. We will take roots or logs -- we will be ready to hang or suspend our rootograms

or rootobars according to chosen comparison curves -- we will use all the skills, tools, and feels we have gained in dealing with counts of amounts and balances.

At the lower level we have to face the two aspects of any count most likely to call for special actions:

- the fact that counts come in units.
- a widespread need for re-expression, usually by roots or logs -- simple or started.

When these two combine, as they usually will, we shall usually find ourselves driven to basic-count intervals of varying length. As a consequence we tend to use rootograms more than rootobars.

27A. Raw counts sometimes work

Some counts of counts deserve to be treated just like counts of amounts or balances. Vansell counted the number of seeds in each pod of eight open-pollinated heads of Ladino clover. Exhibit 1 shows the histograms for the first three heads, exhibit 2 shows the rootograms for the first five heads. As usual, the rootogram seems more effective.

Exhibit 3 contains the calculations for the suspended rootogram for the first head. The only notable feature is the interpolation for the hinges (already familiar to readers of chapter 7x). The essential point is simple -- and one to which we shall recur many times.

exhibit 1 of chapter 27 -- clover pods

Counts of counts;
seeds per pod per head of Ladino clover

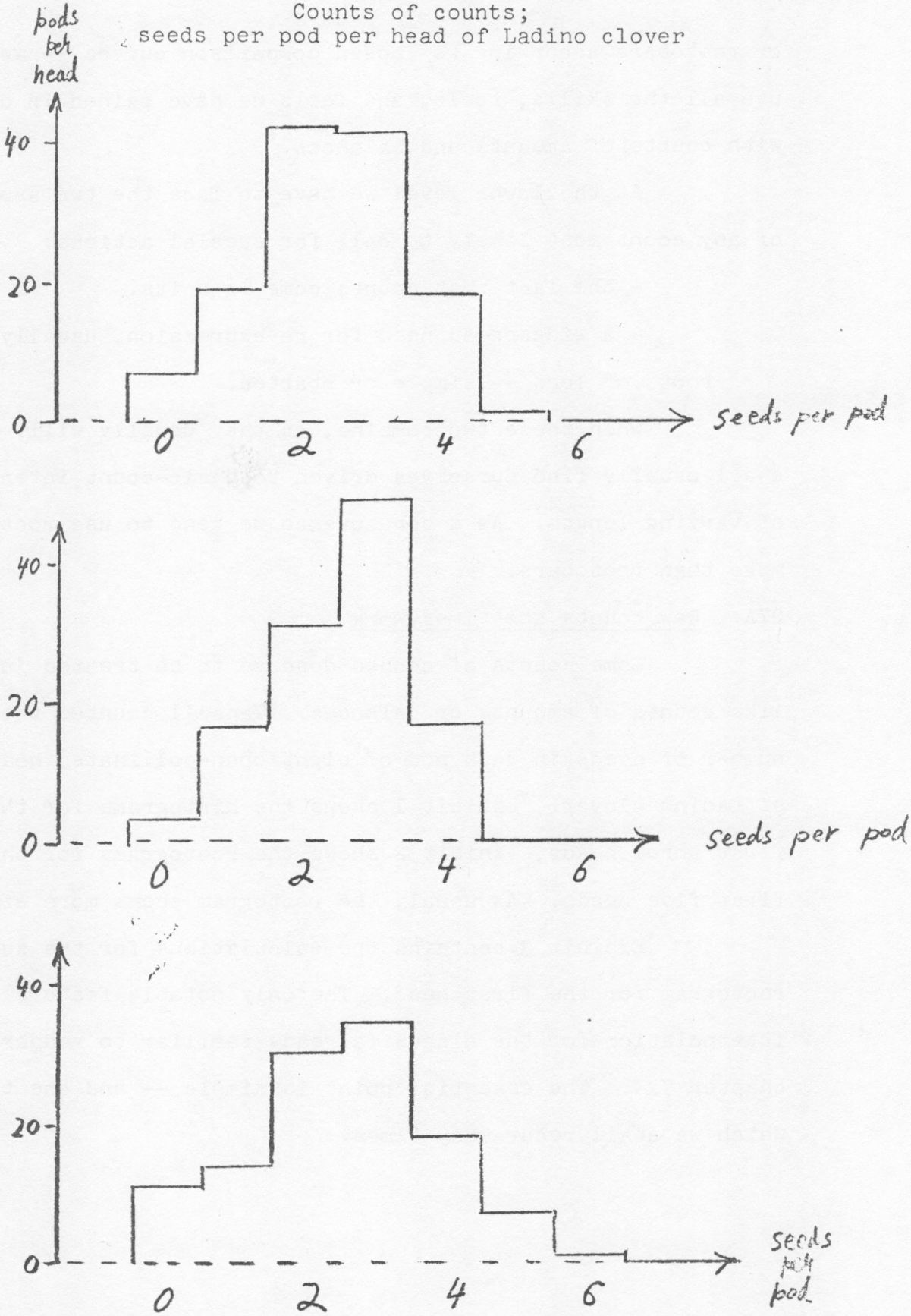


exhibit 2 of chapter 27 -- clover pods

Rootograms for the first five heads

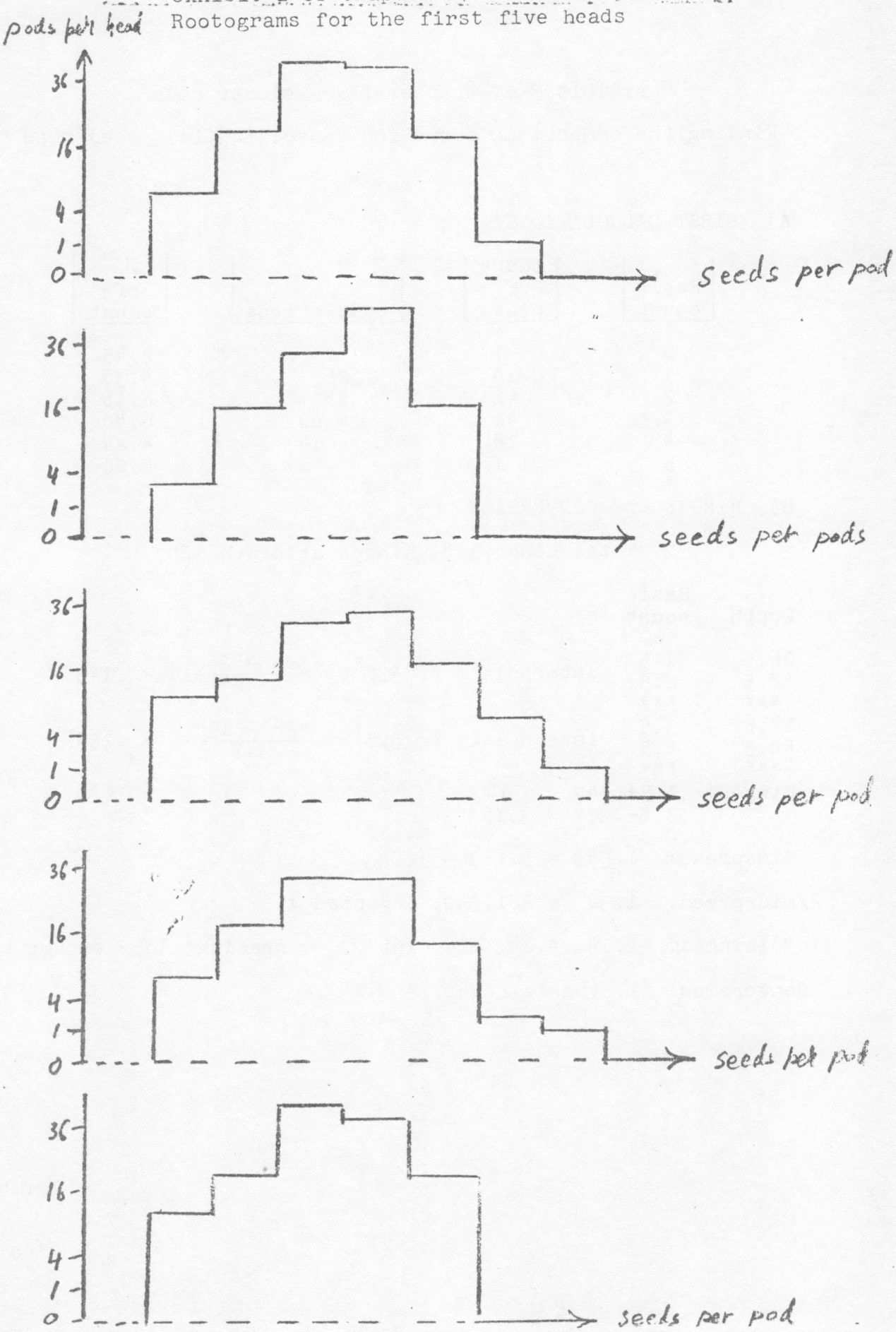


exhibit 3 of chapter 27 -- clover pods

Finding the comparison curve for Vansell's first seed head

A) FIRST CALCULATIONS

Basic count	Count of pods	Cumulations	$\sqrt{}$ of count
0	7	7	2.45
1	19	26	4.36
2	42	68	6.48
3	41	60	6.40
4	18	19	4.24
5	1	1	1.00

B) HINGES and CONVERSION

Total count 128, hinges at depth 32h

Depth	Basic count	
26.5	1.5	interpolate to 32h by $\frac{32h - 26h}{42} \cdot 1 = .143$
68.5	2.5	
***	***	
19.5	3.5	interpolate to 32h by $\frac{32h - 19h}{41} \cdot 1 = .256$
60.5	2.5	
***	***	

Hinges: 1.5+.143 = 1.643
3.5-.317 = 3.183

Midspread: 1.540 = 3.183 - 1.643

2/midspread: 1.30 = 2/1.540, accepted 1.3

(1.3)(hinges): 2.14, 4.14, midpoint 3.14, spacing close enough to 1.00

Conversion: 1.3(basic count) - 3.14

(Cont'd)

exhibit 3 (cont'd.)

C) GAUSS FIT using CONVERSION -- use exhibit 13 of chapter 26, PLEASE

Basic count	$\sqrt{}$ of count	Basic count converted	Relative height for rootogram		Residual
			Tabled	Fitted	
0	2.65	-3.14	.33	2.13	.52
1	4.36	-1.84	.68	4.39	-.03
2	6.48	- .54	.96h	6.23	.25
3	6.40	.76	.93	6.00	.40
4	4.24	2.06	.61	3.94	.30
5	1.00	3.36	.28	1.81	-.81
6	-	4.66	.08h	.55	-.55
		5.96	.01h	.10	-.10
sum	25.13		3.89h	(25.15) ✓	

$$\text{Factor} = \frac{25.13}{3.89h} = 6.46$$

(Dashed)			
-2	-5.3	.04	.25
-1	-4.1	.15	.93

D) SOURCE

G. H. Vansell 1951. Honey bee activity on Ladino Clover florets. J. Econ. Entomol. 44: 103 (only).

An actual basic count corresponds to many ranks or depths -- in the example, basic count 1 corresponds to all depths from 8 to 26. A cut corresponds to one rank or depth -- in the example, basic cut 1.5 corresponds to depth 26.5. To find basic counts corresponding to other depths, we ought to interpolate between basic counts and depths for adjacent cuts. (In emergencies we can do still better, but this is usually good enough.)

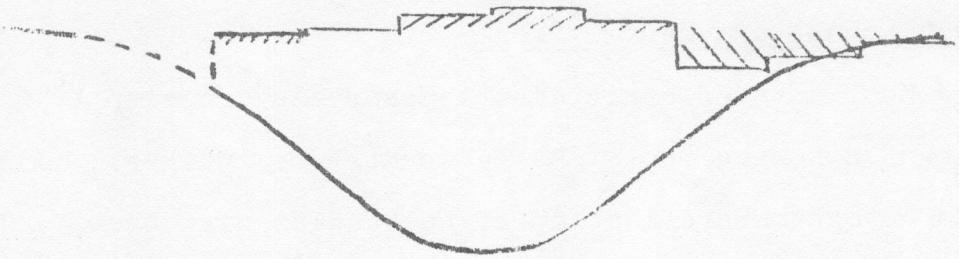
Exhibit 4 shows the suspended rootograms for Vansell's first five seed heads. Clearly there is nothing to gain by changing the expression of these counts in such a way as to squeeze large counts downward.

Raw counts can be useful. (But who can find another example?)

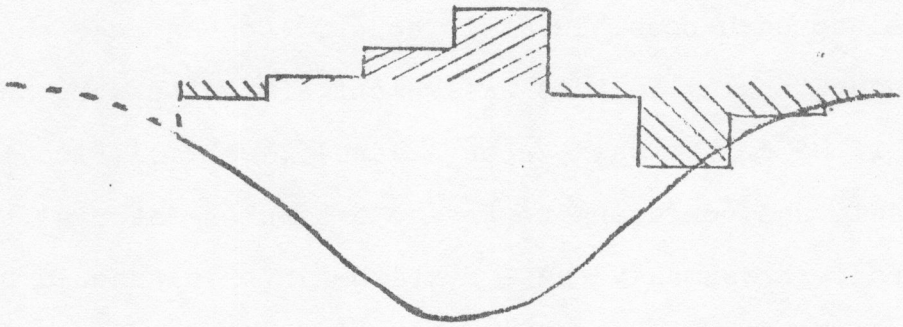
exhibit 4 of chapter 27 -- clover pods

Gauss-suspended rootograms for Vansell's data

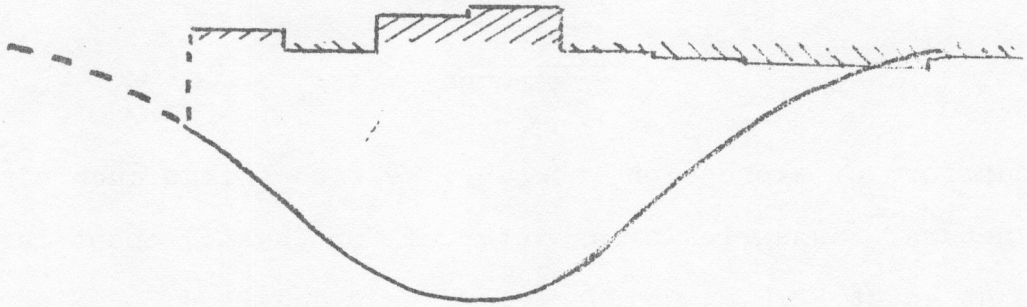
Hinge
(found / used)



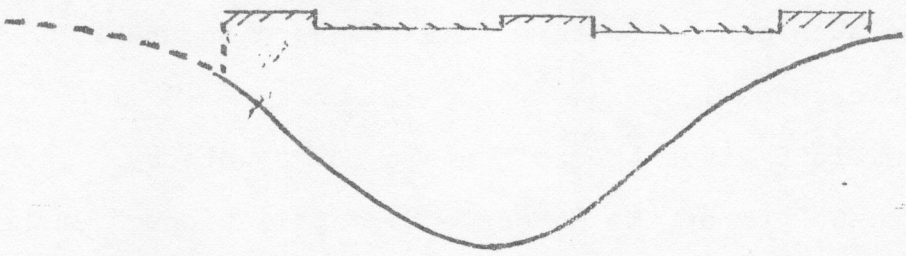
1.64/1.68
3.18/3.18



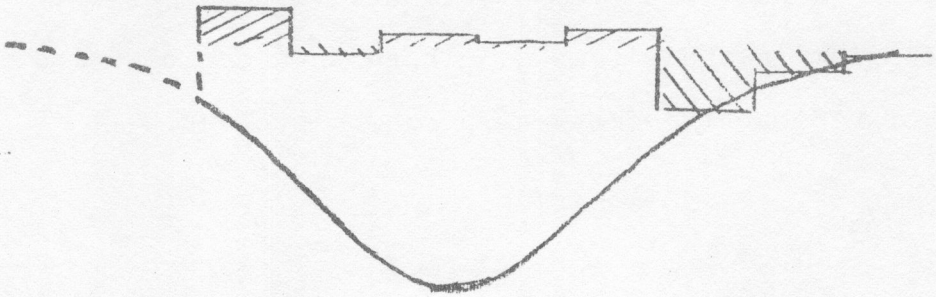
1.70/1.71
3.32/3.31



1.56/1.56
3.46/3.46



1.52/1.52
3.20/3.18



1.56/1.56
3.12/3.12

27B. Root counts are usually best

Most sets of basic counts deserve expression by some sort of a square root. Basic counts are integers. All kinds of rootograms are made up of rectangles with bases. If we are going to make a useful re-expression, we need to assign an interval to each possible basic count. Convenience and ease of understanding in this assignment are quite worthwhile.

If we try to use $\sqrt{\text{count}}$ itself, we will find it uncomfortable and confusing to try to assign an interval to count "zero". Presumably one end of the interval should be at $\sqrt{0.5} = +0.71$, but where should the other end be?

If we decide to use

$$\sqrt{2 + 4(\text{count})}$$

as our form of expression, however, we find things much simpler. The natural boundaries of an interval for (basic) count zero come at count = ±0.5, which gives $\sqrt{0} = 0$ and $\sqrt{4} = 2$. Since other aspects of this choice turn out to be equally desirable, we shall take such double roots as our standard square-root scale for basic counts.

Exhibit 5 gives the arithmetic for counts by Rutherford and Geiger of the number of scintillations (flashes of light) per unit time. Each scintillation indicated the arrival of one α-particle from a radioactive decay of one atom of polonium in their sample. These data are classics in the study of radioactive decay as a random process.

exhibit 5 of chapter 27 -- polonium scintillations

Rutherford and Geiger's data on the radioactive decay of polonium

(Events are scintillations caused by α-particles)

Events per 1/8 minute	Number of of 1/8 minutes	$\sqrt{2 + 4(\text{count})}$ at	cut	Diff	Number per unit	Root of same
			.000			
0	57	1.41		2.000	28h	5.3
			2.000			
1	203	2.45		.828	245	15.6
			2.828			
2	383	3.17		.636	602	24.6
			3.464			
3	525	3.74		.556	980	31.3
			4.000			
4	532	4.24		.472	1128	33.6
			4.472			
5	408	4.69		.427	955	30.9
			4.819			
6	273	5.10		.393	695	26.4
			5.292			
7	139	5.47		.365	381	19.5
			5.657			
8	45	5.83		.343	131	11.4
			6.000			
9	27	6.16		.325	83	9.1
			6.325			
10	10	6.48		.308	32	5.7
			6.633			
11	4	6.78		.295	13h	3.7
			6.928			
12	0	7.07		.283	0	0
			7.211			
13	0	7.35		.272	0	0
			7.483			
14	1	7.62		.263	3.8	2.0
			7.746			
(sum 1)	(2607)					(219.1)

Note: Ratios and roots to slide rule accuracy.

Examples: $\begin{cases} 0.000 = \sqrt{0},\ 1.41 = \sqrt{2},\ 2.000 = \sqrt{4},\ 2.45 = \sqrt{6},\ 2.828 = \sqrt{8},\ldots \\ 2.000 = 2.000-0.000,\ .828 = 2.828-2.000,\ldots \end{cases}$

$28.5 = 57/2.000,\ 245 = 203/.828,\ldots$

$5.3 = \sqrt{28.5},\ 15.6 = \sqrt{245},\ldots$

Source: E. Rutherford and H. Geiger 1910. The probability variations in the distribution of α-particles. Philosophical Magazine 10: 698-704. Table on page 701.

Exhibit 6 shows the constant-base histogram for this data. (The first two columns of exhibit 5 provide all we need to plot it.) If we needed to be told that the wall at zero keeps small counts from straying downward as far as they stray upward, we could make good use of this plot -- it can tell us this. But we learned this long ago. It is hard to learn much more from exhibit 6.

Exhibit 7 shows the variable-base histogram corresponding to our standard square-root scale for basic counts. (The "number per unit" column of exhibit 5 gives us the needed heights). It is reasonably symmetric -- we are inclined to believe that we might see an anomaly if one were present. The tails, as always with histograms, look small and might well hide anomalies.

Exhibit 8 shows us the corresponding rootogram. (The "Root of same" column in exhibit 5 now gives us the needed heights.) As usual, we have all the advantages of exhibit 7 except area-trueness, and we have a very much better look at the tails. This is a further improvement.

Exhibit 9 shows the same information in rootobar form. Some will find it an improvement. Most will find it "busy" and notably less satisfactory.

Some will find exhibit 10 -- a simplified and skeletonized form of exhibit 8 -- an improvement on the ordinary rootogram, others will not.

Clearly it is time to try suspending our rootogram.

exhibit 6 of chapter 27 -- polonium scintillations

Rutherford and Geiger's data
(Worst display, histogram)

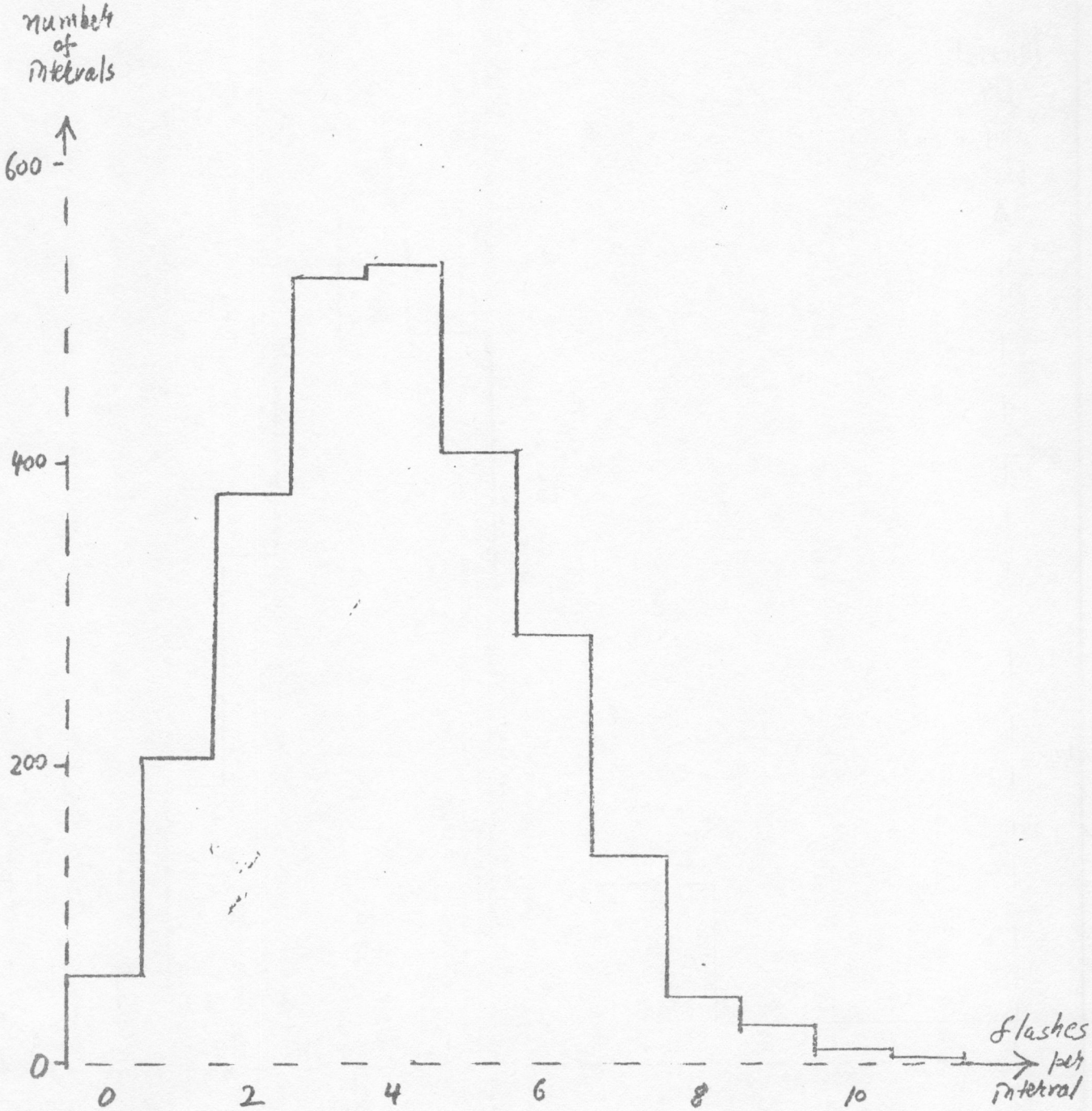


exhibit 7 of chapter 27 -- polonium scintillations

Rutherford and Geiger's data again
(better display, double-root-base histogram)

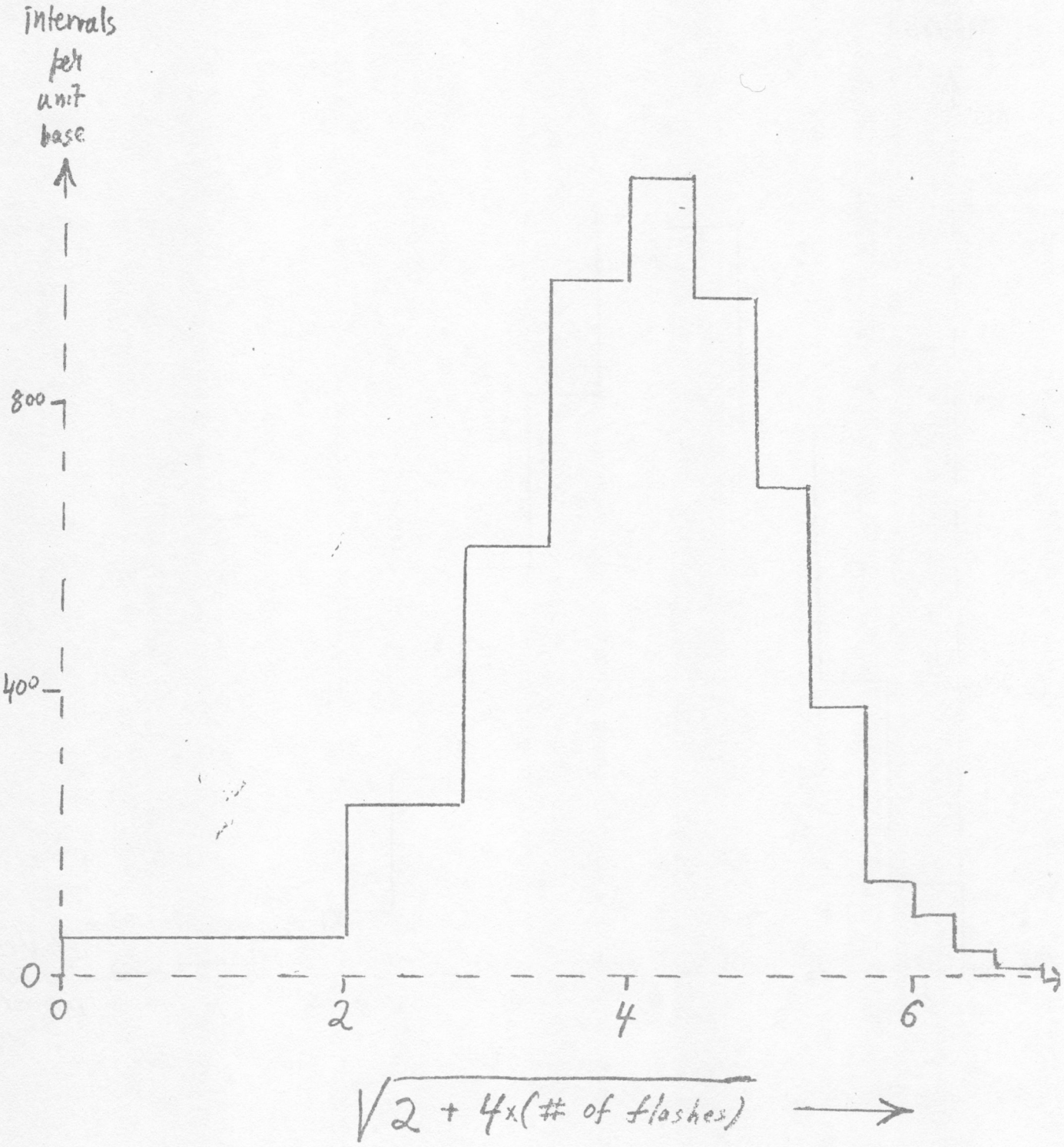


exhibit 8 of chapter 27 -- polonium scintillations

Rutherford and Geiger's data yet again
(still better display,
double-root-base rootogram)

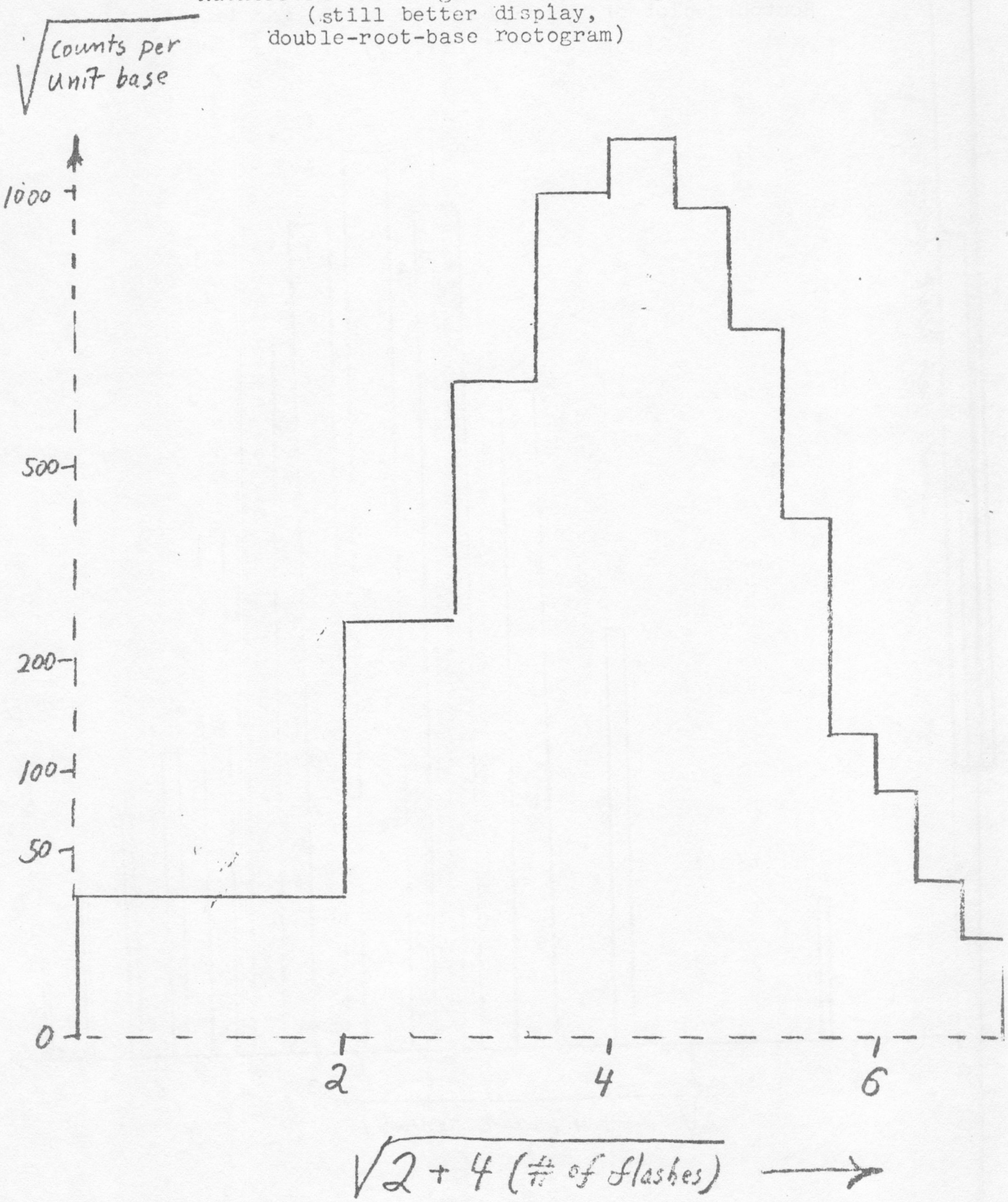


exhibit 9 of chapter 27 -- polonium scintillations

Rootobar plot of the Rutherford and Geiger data
(not too satisfactory)

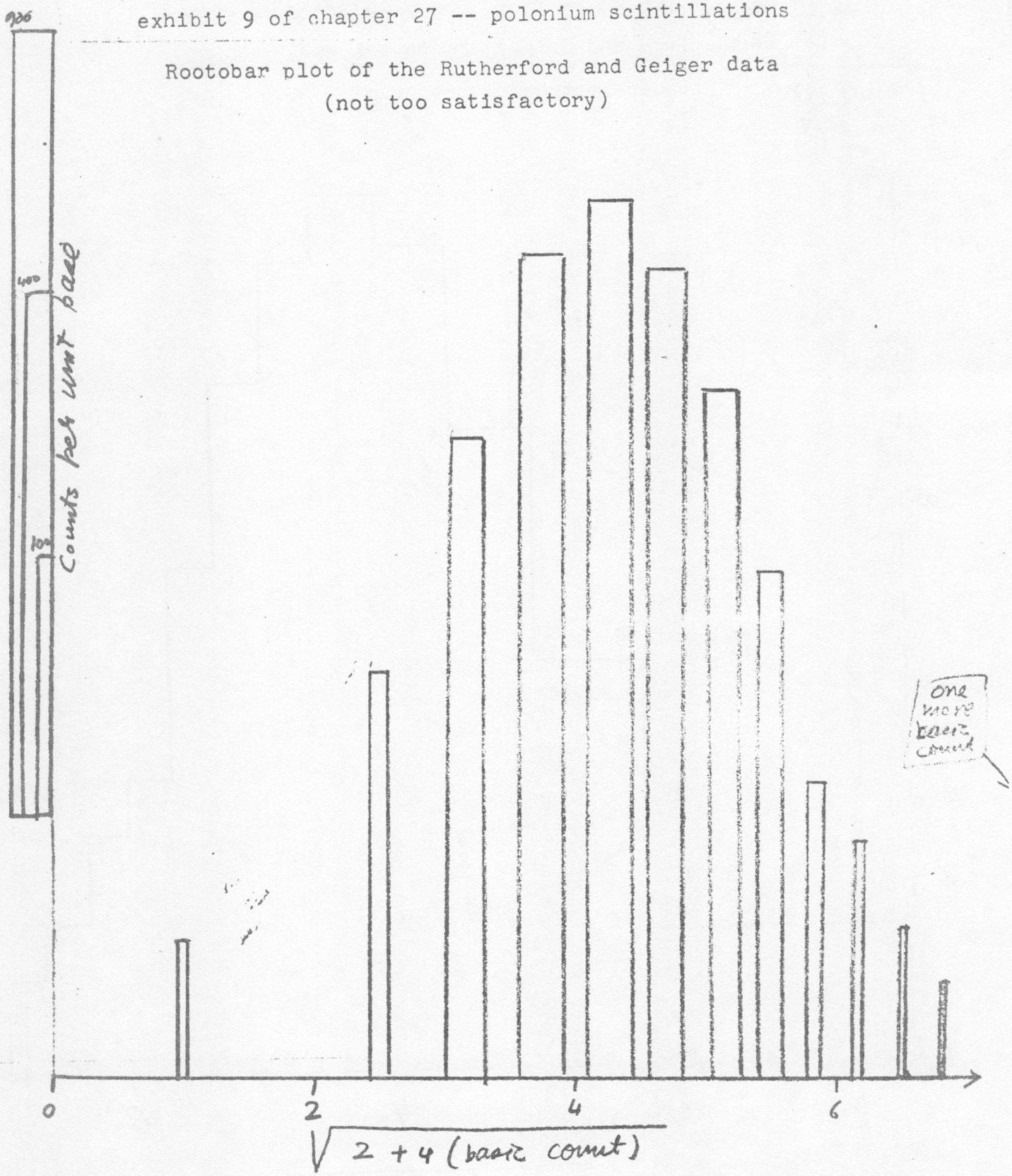
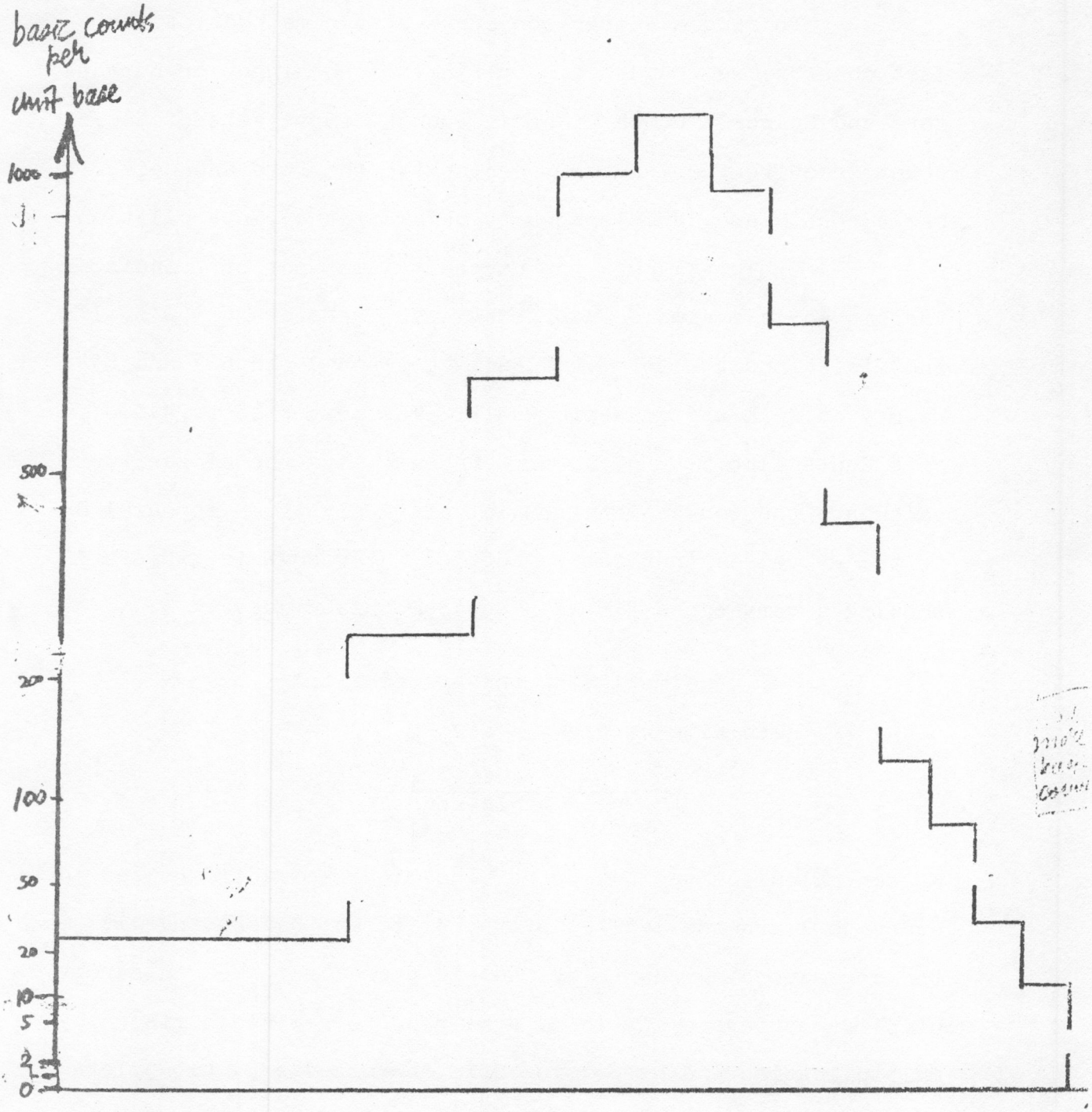


exhibit 10 of chapter 27 -- polonium scintillations

Skeletonized form of rootogram



* the comparison curve *

To find a comparison curve by the methods of the last chapter, we would begin by finding the hinges of Rutherford and Geiger's batch of basic counts. Some sets of basic counts require special care here, but Rutherford and Geiger's basic counts are large enough to be treated by interpolation.

Exhibit 11 has the necessary arithmetic. Finding the hinges is a simple interpolation. There are 643 0's, 1's, and 2's. Thus a $\sqrt{\ }$ of 3.464 falls between depth 643 and 644. We may as well assign depth 643.5. The next $\sqrt{\ }$ cut, 4.000, gets depth 1168.5. And so on. Once we have the hinges, we calculate the conversion formula, and apply it -- in panel B -- to the $\sqrt{\ }$ values (not the $\sqrt{\ }$ cuts). The result is exhibit 12, showing a very close fit with the comparison curve.

* simplifying the arithmetic *

Each time we use

$$\sqrt{2 + 4(\text{basic count})}$$

we can calculate the cut points and find their differences in order to learn the lengths of the different bases, and divide the corresponding counts by these differences. We can surely simplify matters a little by arranging to replace division by multiplication and by putting down the necessary multipliers in a table. Exhibit 13 starts to do this, and notices that the multipliers are simply related to the values of $\sqrt{2 + 4(\text{basic count})}$ themselves.

exhibit 11 of chapter 27 -- polonium scintillations

Computations for suspending the rootogram
from a Gaussian active

A) HINGES and CONVERSION FORMULA

Basic count	Count	Depth of cut	$\sqrt{\ }$ of cut	Hinge	1.6 times same	less 6.557
0	57					
1	263	57				
2	383	260				
3	525	643	3.464			
4	532	1168	4.000	3.473	5.557	-1.00
5	408	907	4.472			
6	273	499	4.899	4.723	7.557	+1.00
7	139	226				
8	45	87				
9	27	42				
10	10	15				
11	4	5				
12	0	1				
13	0	1				
14	1	1				
15						

Depth of hinges = 2607/4 → 652

Interpolation: $\frac{8.5}{525}$ (.536) = .009, 3.464+.009 = 3.473

$\frac{152.5}{408}$ (.472) = .176, 4.899-.176 = 4.723

1.250 = 4.723 - 3.473, .625 = 1.250/2

1/.625 = 1.6

Converted value = 1.6($\sqrt{\ }$) - 6.56

(Cont'd)

exhibit 11 of chapter 27 (cont'd)

B) FIXING the COMPARISON CURVE

Basic count	$\sqrt{\ }$	Conversion	Rootogram comparison height: Tabled*	Rootogram comparison height: Matched	Obs'd	Diff
0	1.41	-4.30	.12	3.9	5.3	1.4
1	2.45	-2.64	.45	14.7	15.6	0.9
2	3.17	-1.49	.77	25.2	24.6	-0.6
3	3.74	-.57	.96	31.4	31.3	-0.1
4	4.24	.23	.99	32.4	33.6	1.2
5	4.69	.94	.90	29.5	30.9	1.4
6	5.10	1.59	.75	24.6	26.4	1.8
7	5.47	2.19	.57	18.7	19.5	0.8
8	5.83	2.76	.42	13.8	11.4	-2.4
9	6.16	3.30	.29	9.5	9.1	-0.4
10	6.48	3.82	.19	6.2	5.7	-0.5
11	6.78	4.32	.12	3.9	3.7	-0.2
12	7.07	4.76	.07	2.3	0	-2.3
13	7.35	5.22	.04	1.3	0	-1.3
14	7.62	5.64	.03	1.0	2.0	1.0
15	7.87	6.03	.01	0.3	0	
			(sum)(6.70)		(219.1)	
Extra	0	-6.56	.01	0.3		
	1.00	-4.96	.06	2.0		
	4.10	0	1.00	32.8		
			.01	0.3		

Examples: $-4.30 = 1.6(1.41) - 6.56$, $-2.64 = 1.6(1.41) - 6.56$

$$\text{factor} = \frac{219.1}{6.70} = 32.7$$

* From exhibit 13, chapter 26, NOT from exhibit 10).

exhibit 12 of chapter 27 -- polonium scintillations

The Gauss-suspended rootogram
(Based on exhibit 11)

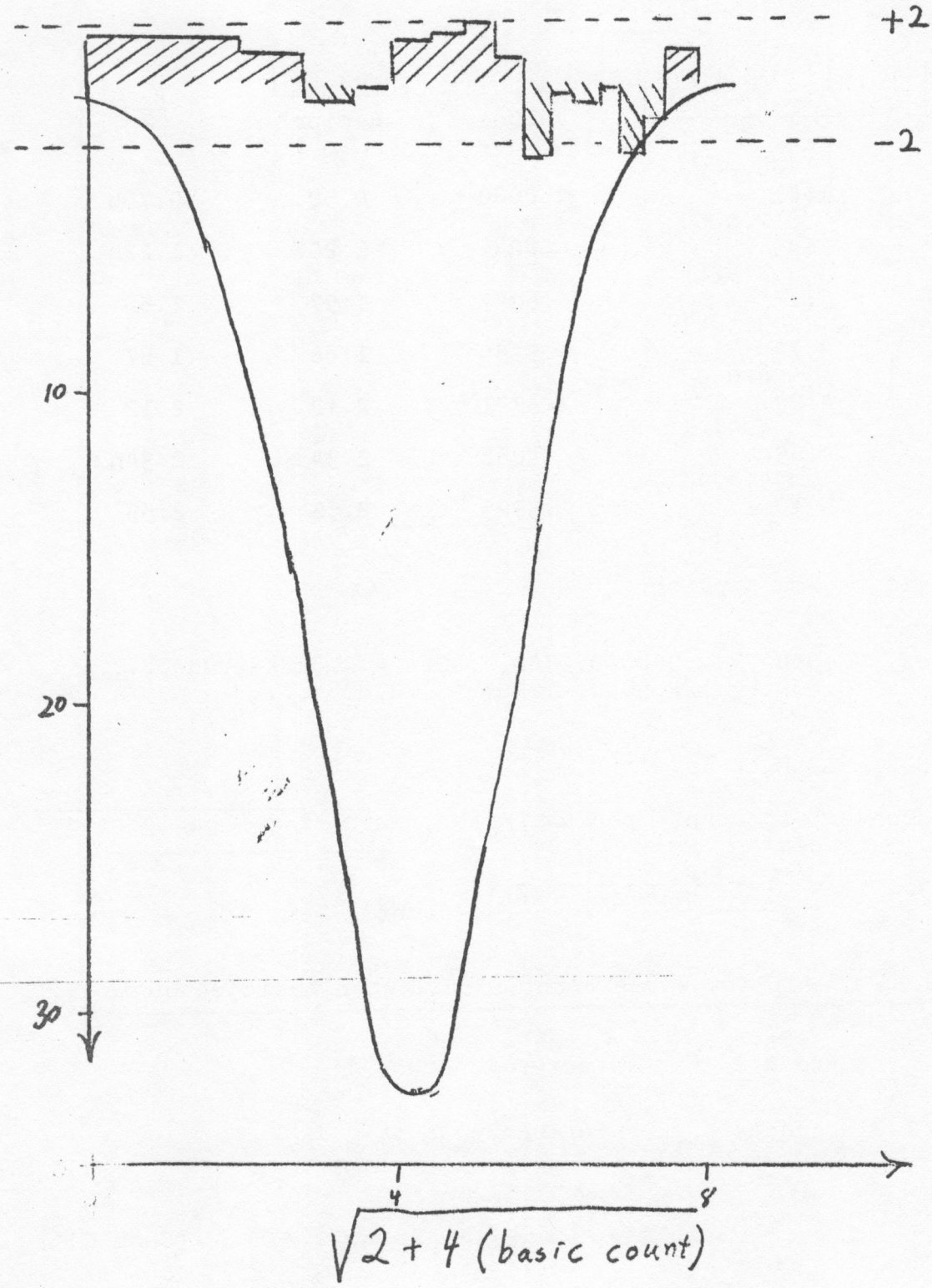


exhibit 13 of chapter 27 -- doubled-root spacings

Streamlining the calculations

A) LISTING the MULTIPLIERS

Basic count	$\sqrt{2 + 4(\text{same})}$	Differences: Values	Differences: Reciprr*	$\frac{\sqrt{\ }}{2}$
-0.5	0.0000			
0	1.41	2.0000	0.50	0.70h
0.5	2.0000			
1	2.45	.8284	1.207	1.22h
1.5	2.8284			
2	3.17	.6357	1.57	1.58
2.5	3.4641			
3	3.74	.5359	1.86	1.87
3.5	4.0000			
4	4.24	.4721	2.12	2.12
4.5	4.4721			
5	4.69	.4268	2.34	2.34h
5.5	4.8990			
6	5.10	.3925	2.55	2.55
6.5	5.2915			

* This is the multiplier.

Examples: 2.0000 = 2.0000-0.0000, .8284 = 2.8284-2.0000,...
0.500 = 1/2.0000, 1.207 = 1/0.8284,...

B) THE SIMPLE RULE

To convert to counts per unit base, when $\sqrt{2 + 4(\text{basic count})}$ is used, calculate

$$\frac{\sqrt{2 + 4(\text{basic count})}}{2} \text{ (count)}$$

and use as multiplier times count EXCEPT when basic count is 0 or 1. There use

$\frac{1}{2}$ (count) for basic count = 0

1.21 (count) for basic count = 1

(Cont'd)

exhibit 13 (Cont'd)

C) WORKED EXAMPLE -- yeast in haemacytometer.

Basic count	√	Count	(1/2)* Product	Cut
				0.00
0	1.41	103	51.5	
				2.00
1	2.45	202	244.4	
				2.82
2	3.17	268	426.	
				3.46
3	3.74	143	268.	
				4.00
4	4.24	32	68.0	
				4.47
5	4.69	17	39.9	
				4.90
6	5.10	10	25.5	
				5.29
7	5.58	0	0	

* Except for basic counts 0 and 1, one-half of "√" times "count".

D) SOURCE of DATA in panel C.

Student 1907. On the error of counting with a haemacytometer. Biometrika: 5: 351-360. Example (ii) on page 356.

Panel C illustrates the short computation for one of Student's famous counts of the numbers of yeast cells on the squares of a hemocytometer -- a device ruled into 400 squares in a 20x20 pattern that is used for counting blood cells. Exhibit 14 shows the rootogram

27C. Further examples

Let us recall exhibit 4 of chapter 24, where we looked at Ashby's counts of Salicornia europea plants in 625 square-centimeter squares. There we found roots the more reasonable choice. Now that we have our technique sharpened, let us find and plot the suspended rootogram for this data. Intervals of 10 in the count -- used in panel B of exhibit 4 of chapter 24 -- seem likely to give a helpful picture. Hinges can be calculated without interpolation. They come out as 20 and 68h. Their double roots are 9.06 and 16.55. The conversion is thus .267(double root) - 3.42.

Exhibit 15 shows the streamlined calculations for rootograms. Exhibit 16 shows the plain rootogram. Clearly it gives us a reasonable look at the pattern of counts. We would expect to be able to see anomalies were they at all substantial. We would also expect to be able to compare patterns for different sets of counts.

Exhibit 17 shows the suspended rootogram. The observations seem to deviate systematically from the comparison curve, but it would clearly take more than 100 counts to give us a secure feeling that this was not an accident.

exhibit 14 of chapter 27 -- yeast in haemacytometer

One set of Student's counts of yeast cells in a haemacytometer
(see panel C of exhibit 13)

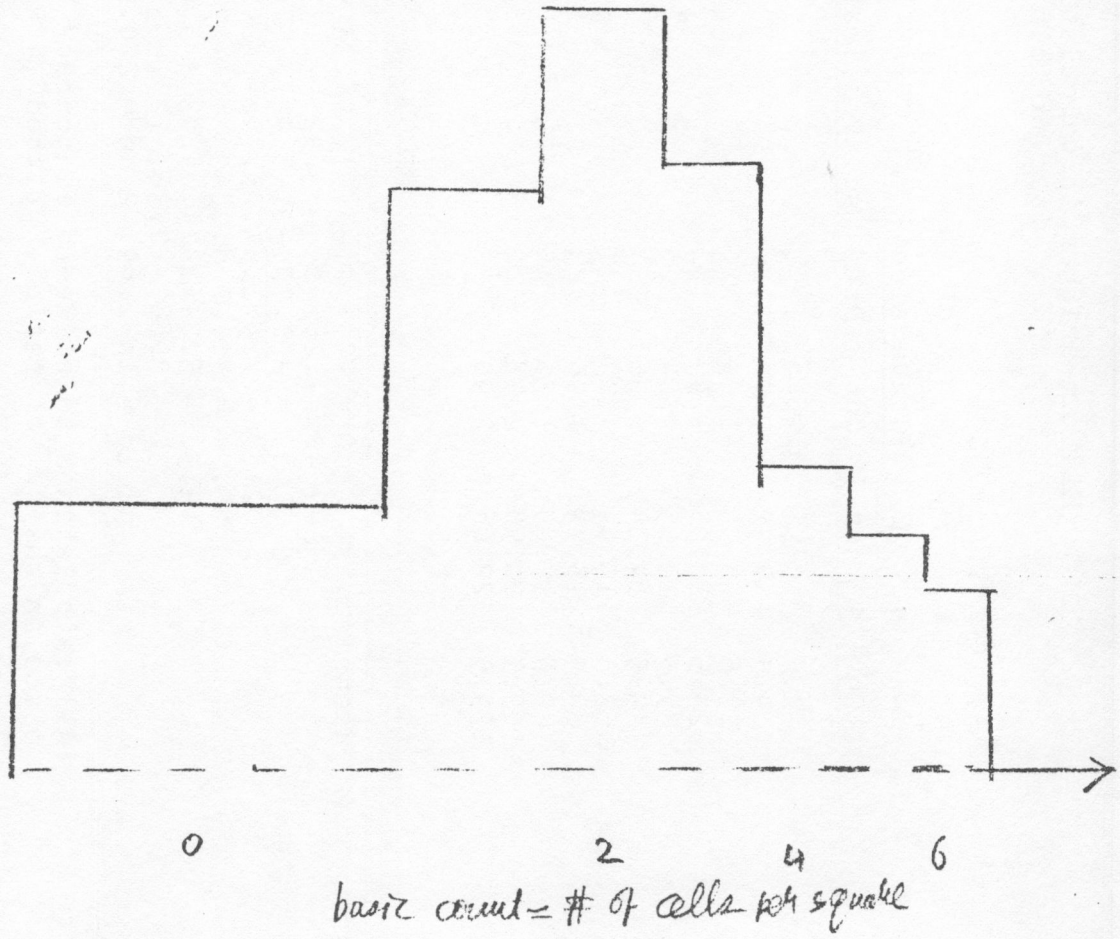


exhibit 15 of chapter 27 -- Salicornia quadrats

Minimum-arithmetic calculation of rootograms for Ashby's data
(see exhibit 4 of chapter 24)

A) CALCULATIONS

Cut	Double root	Conversion: Cut	Conversion: Diff	count of squares	√ of ratio*	Gauss roots: Table	Gauss roots: Fitted	Residuals	±1 in √#**	±1h in √#
-0.5	0	-3.42								
			1.69	14	2.88	.47	2.36	.52	± .77	±1.15
9.5	6.33	-1.73								
			.70	10	3.78	.80	4.03	-.25	±1.20	±1.80
19.5	8.94	-1.03								
			.54	12	4.72	.93	4.68	.04	±1.36	±2.04
29.5	10.95	-.49								
			.45	8	4.22	.99	4.98	-.76	±1.49	±2.24
39.5	12.65	-.04								
			.40	8	4.47	.99h	5.01	-.54	±1.58	±2.37
49.5	14.14	.36								
			.36	12	5.76	.96h	4.84	.92	±1.67	±2.51
59.5	15.49	.72								
			.34	12	5.95	.91	4.58	1.37	±1.71	±2.56
69.5	16.73	1.06								
			.30	11	6.05	.84h	4.26	1.79	±1.83	±2.75
79.5	17.89	1.36								
			.82	8	3.13	.70	3.52	-.39	±1.10	±1.65
109.5	20.98	2.18								
			.72	5	2.63	.48	2.42	.21	±1.18	±1.77
139.5	23.66	2.90								
(sum)				(100)	(40.71)	(8.08h)	(40.68√)			

Examples:

$0 = \sqrt{2 + 4(-0.5)}$, $6.33 = \sqrt{2 + 4(9.5)}$,...
$-3.42 = .267(0) - 3.42$, $-1.73 = .267(6.33) - 3.42$,...
$1.69 = -1.73 - (-3.42)$, $.70 = -1.03 - (-1.73)$,...
$2.88 = \sqrt{14/1.69}$, $5.84 = \sqrt{24/.70}$
$5.04 = 40.71/8.08h$
$2.36 = 5.04(.47)$, $4.03 = 5.04(.80)$,...
$.52 = 2.88 - 2.36$, $.25 = 3.78 - 4.03$,...
$1.18 = 1h/1.69$, $2.86 = 2/0.70$,...

All calculations to slide-rule accuracy only. (We don't really need the second decimal place in the "√ of ratio", "fitted", "Residual", and "2 in √" columns.)

To plot rootogram do a staircase on **cut** and "√ of product" columns.

To plot suspended rootogram, use "fitted" and "residuals" between first column "cut".

* This column is $\sqrt{(\text{count})/\text{diff}}$

** This column is $(\pm 1)/\sqrt{\text{diff}}$, the next is 1h times this.

(cont'd.)

exhibit 15 of chapter 27 (cont'd.)

B) SOURCE

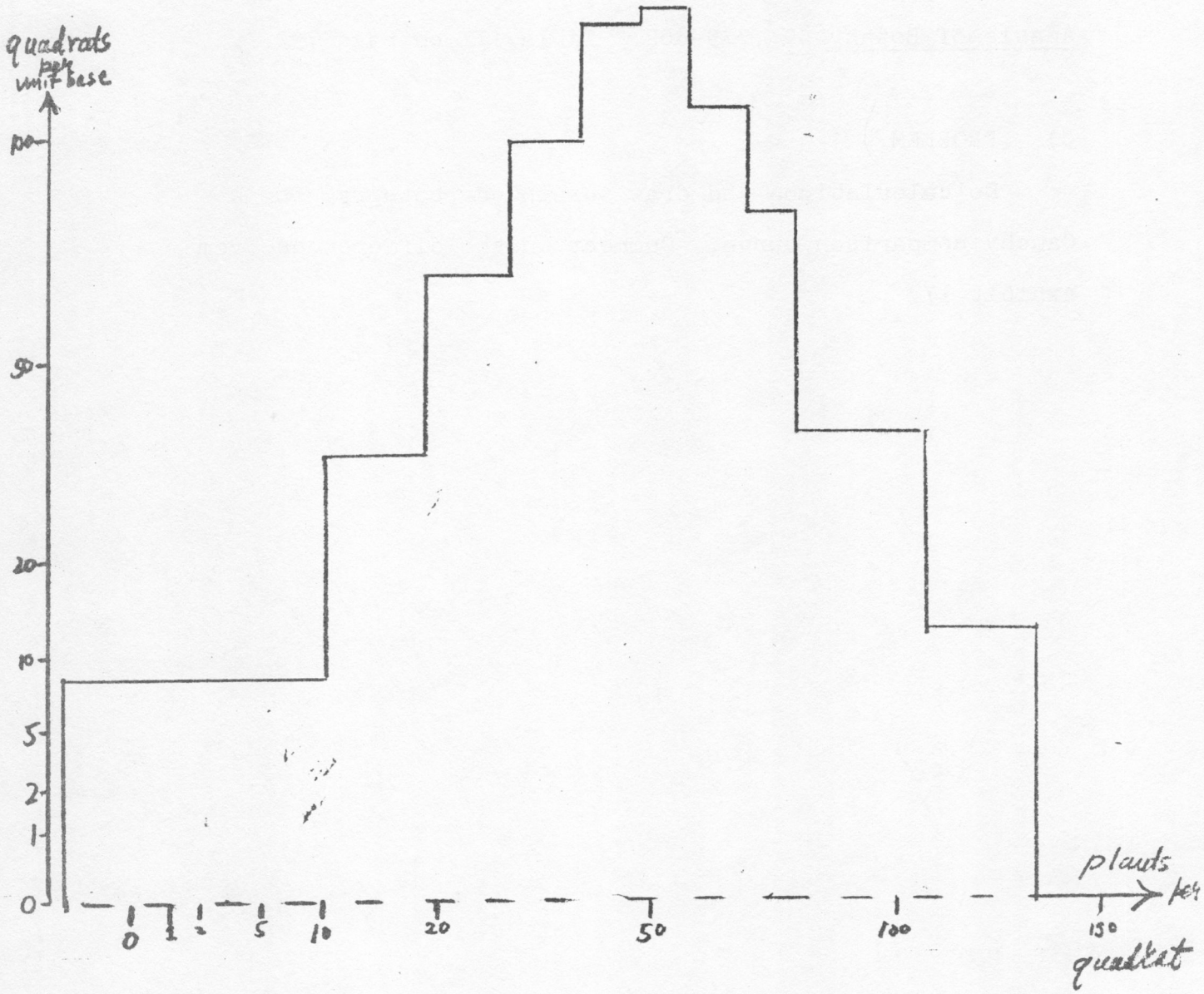
E. Ashby 1935. The quantitative analysis of vegetation. Annals of Botany 49: 779-802. Table III on page 790.

C) PROBLEM

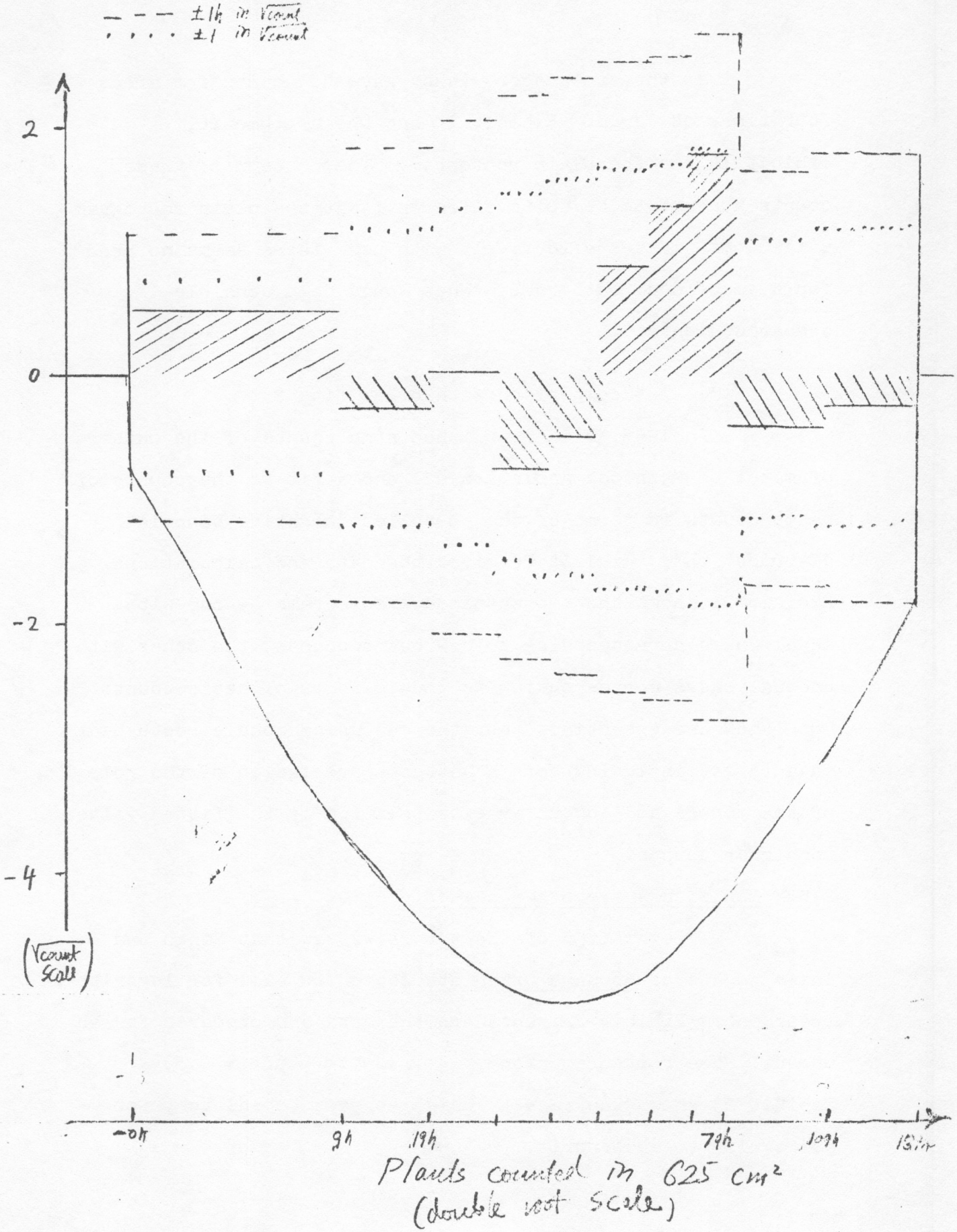
Do calculations and draw suspended rootogram for a Cauchy comparison curve. Comment on the differences from exhibit 17.

exhibit 16 of chapter 27 -- Salicornia quadrats

Ashby's counts of Salicornia europea



Gauss-suspended rootogram for Ashby's counts
(see exhibits 15 and 16)



In the same paper, Ashby gave 58 counts for areas four times as large. Exhibit 18 has the arithmetic, exhibit 19 has the plain rootogram. The pattern of these counts appears to be quite different; but the plain rootogram still communicates effectively with us. There seems no need for a suspended rootogram. (What would be a sensible comparison curve?)

* comparison with other fits *

In 1953, Bliss published some counts of the numbers of mites on McIntosh apple leaves, and a fit to the counts of basic counts in terms of the so-called "negative binomial distribution". Exhibit 20 gives the data and calculations. Exhibit 21 shows the two suspended rootograms -- one with equal bases corresponding to raw basic counts, the other with unequal bases corresponding to double roots of basic counts. Both show the essentials, but the one using double roots has advantages, including both a better appreciation of the role of the zeroes and an easier extrapolation of the fitted values to higher counts.

27D. Log counts are often needed

In exhibit 5 of chapter 24 we saw that Singh and Das's counts of Bornaya brachiata seemed to call for logarithmic treatment. Exhibit 22 considers the data and prepares for an unequal base rootogram using log (basic count + 0.5) . Exhibit 23 shows the result, which appears too blotchy for us to wish to suspend it.

exhibit 18 of chapter 27 -- larger quadrats

Ashby's counts of Salicornia europea for larger squares
(2500 cm^2 in area)

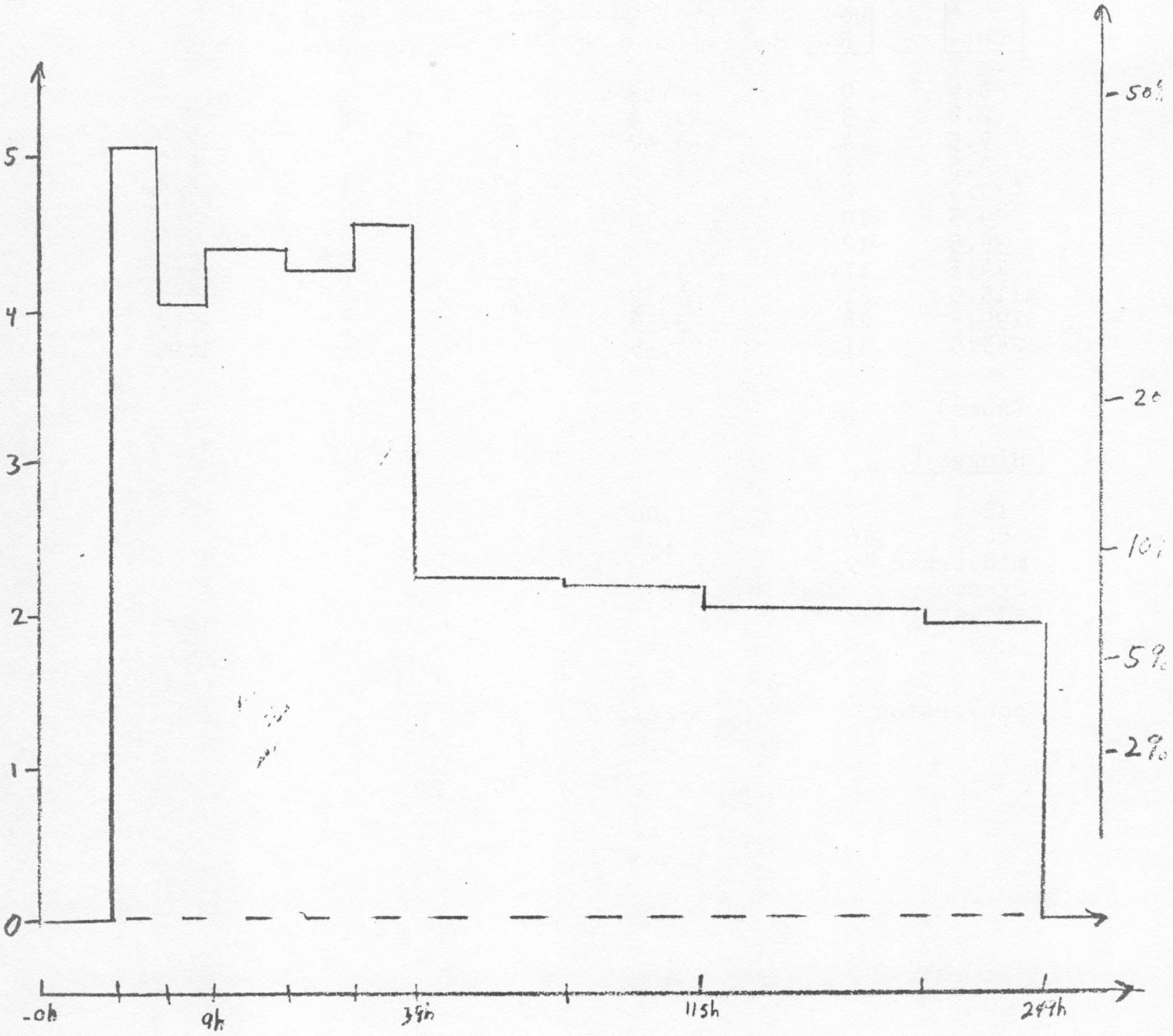
Cut	Double Root	Conversion Cut	Conversion Diff	Count of squares	√ of product
-0.5	0	-2.53			
2.5	3.46	-1.78	.75	0	0
5.5	4.90	-1.47	.31	8	5.09
9.5	6.32	-1.16	.31	5	4.02
19.5	8.94	-.59	.57	11	4.40
29.5	10.95	-.16	.43	8	4.30
39.5	12.65	.21	.37	8	4.60
74.5	17.32	1.21	1.00	5	2.24
115.5	21.54	2.13	.92	4	2.08
199.5	28.28	3.58	1.45	6	2.04
249.5	31.62	4.35	.77	3	1.96
(sums)				(58)	

Hinges		
12	7.07	-1.00
66	16.31	+1.00
midspread	9.24	
2/same	.216	
adjust	-2.53	

conversion: $-2.53 + .216\sqrt{2 + 4\times(\text{count})}$

exhibit 19 of chapter 27 -- larger quadrats

Ashby's counts of Salicornia europea for larger squares
(see exhibit 18)



Plants counted in 2500 cm²
(double root scale)

exhibit 20 of chapter 27 -- apple mites

Garman's counts of mites on McIntosh apple leaves

A) DATA and EQUAL-BASE CALCULATIONS

No. of found mites	Obs'd no. of leaves	√ of obs'd	fitted no. of leaves	√ of fitted	Residual
0	70	8.36	69.5	8.34	.02
1	38	6.16	37.6	6.13	.03
2	17	4.12	20.1	4.48	-.32
3	10	3.16	10.7	3.27	-.11
4	9	3.00	5.7	2.39	.61
5	3	1.73	3.0	1.73	.00
6	2	1.41	1.6	1.26	.15
7	1	1.00	.85	.92	.08
≥8	(0)	(0.00)	(.95)	(.97)	(-.97)

B) UNEQUAL-BASE CALCULATIONS

	Double root		√counts/unit base			
Cut	Cut	Diff	Obs'd	Fitted	Residual	(±1)
-0.5	0					
		2.00	5.93	5.90	.03	(± .5)
0.5	2.00					
		.83	6.76	6.73	.03	(±1.2)
1.5	2.83					
		.63	5.20	5.67	-.47	(±1.6)
2.5	3.46					
		.54	4.31	4.44	-.13	(±1.8)
3.5	4.00					
		.47	4.38	3.47	.91	(±2.1)
4.5	4.47					
		.43	2.64	2.64	.00	(±2.3)
5.5	4.90					
		.39	2.27	2.02	.00	(±2.6)
6.5	5.29					
		.36	1.67	1.54	.13	(±2.7)
7.5	5.65					
		.35	0	?	?	(±2.8)
8.5	6.00					

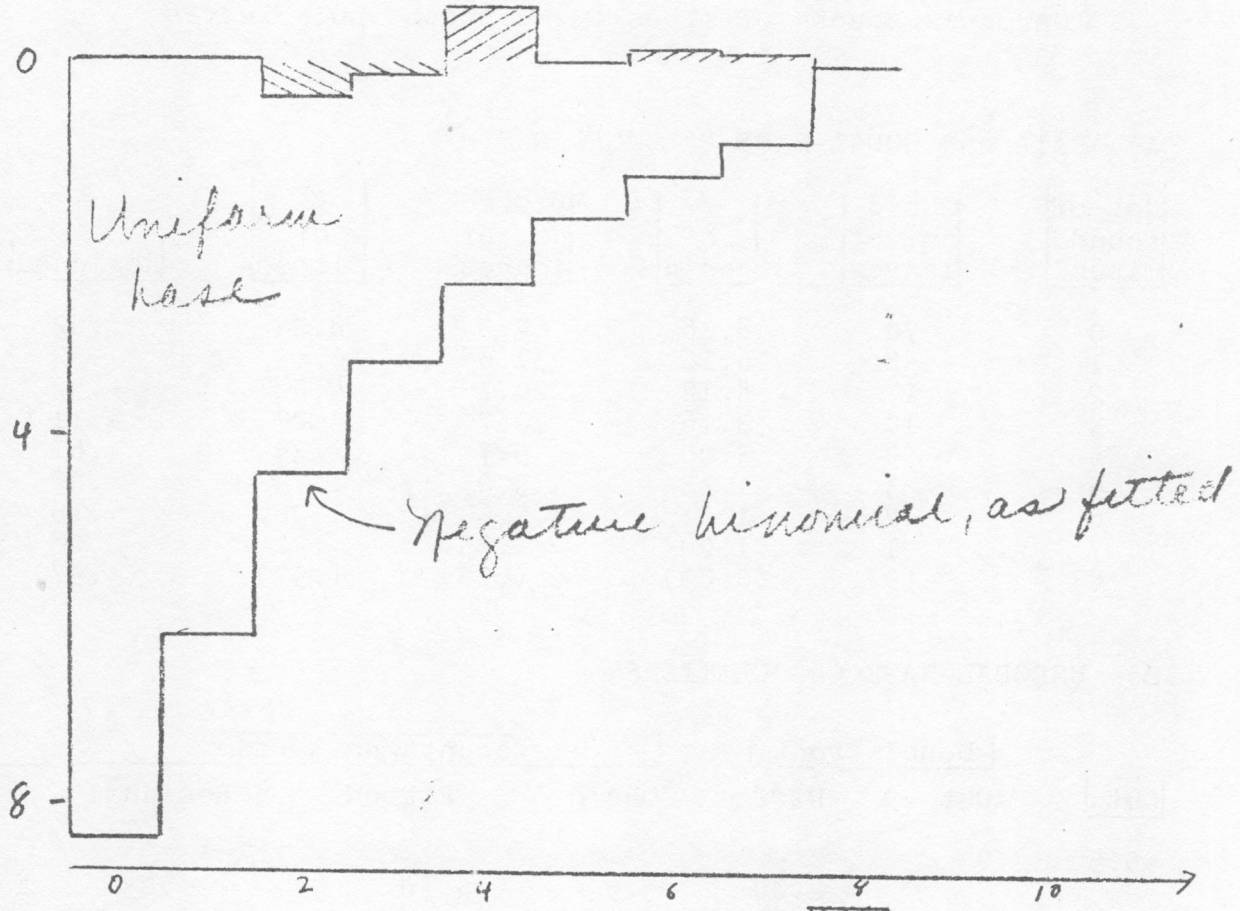
* ±1 in √count

C) SOURCE

C. I. Bliss 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176-190. Table 1 on page 178. His source: Unpublished data of P. Garman.

Garman's counts of mites on McIntosh apple leaves

(see exhibit 20)



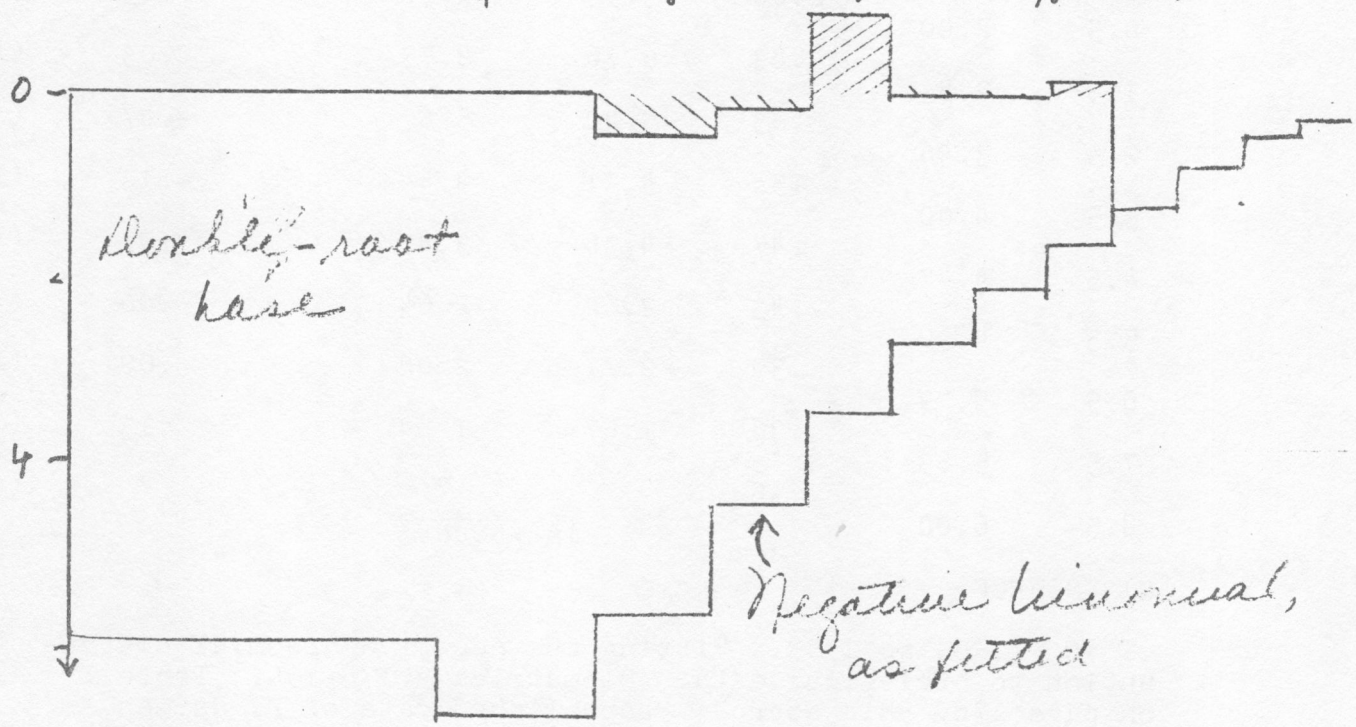


exhibit 22 of chapter 27 -- Bornaya infestation

Calculations for a rootogram for log (0.5 + basic count)
for counts of Bornaya brachiata
(see exhibit 5 of chapter 24)

A) DATA CALCULATIONS

Cut	log(0.5+count) Cut	Diff	Count of plants	Root of count/diff
3.5	0.60			
		.30	13	6.57
7.5	0.90			
		.30	13	6.51
15.5	1.20			
		.30*	26	9.30
31.5	1.51			
		.30	20	8.18
63.5	1.81			
		.30	16	7.31
127.5	2.11			
		.30	12	6.30
255.5	2.41			

* All differences are .301 when calculations are taken to enough decimals, so we can neglect the apparent .31.

B) SOURCE

B. N. Singh and K. Das 1938. Distribution of weed species on arable land. Journal of Ecology 26:455-466. Table II on page 461.

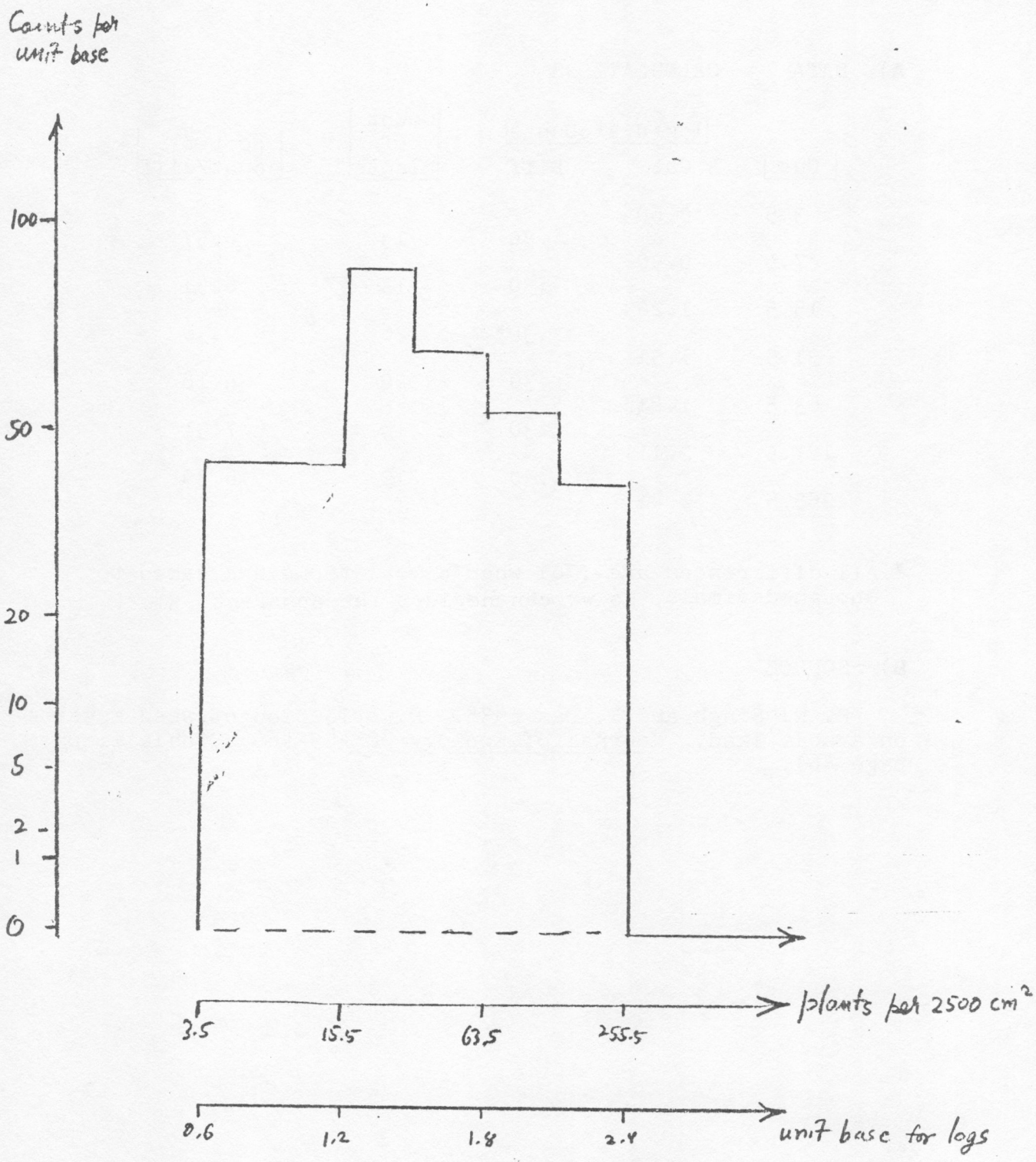
exhibit 23 of chapter 27 -- Bornaya infestation

Rootogram of

log (0.5 + plants per 2500 cm^2)

for Bornaya brachiata

(see exhibit 22)



Note the convenience of using log (0.5 + basic count) instead of an expression with some other free start. The cuts come at half-integers. Thus 0.5 + cut comes at integers, which we can often take -- as in this example -- to be powers of 2. This converts what would otherwise be an unequal-base calculation into an equal base one, simplifying our arithmetic.

In 1948, Preston published counts of the numbers of breeding pairs of different species of birds in 16,967 acres of Quaker Run Valley. It is easy to see that these numbers need to be looked at in terms of logarithms. Exhibit 24 gives the data and the calculation for rootograms of log (0.5 + number of pairs). The plain rootogram -- given in exhibit 25 -- has a reasonable shape, so we ask to see the suspended rootogram.

Exhibit 26 shows it to us. The fit by the comparison curve is relatively close, the only noticeable deviations being the absence of species with more than 2048 breeding pairs where about three would be expected from the fit. Clearly modifying the free start in log (0.5 + number of pairs) -- which will matter only for small numbers of pairs -- is not called for.

Although this example deals with "species", the methods more often used for counts of instances work very well -- better than the special methods for species, as we shall see in chapter 27x.

exhibit 24 of chapter 27 -- breeding birds

Numbers of breeding pairs of birds by species in 16,967 acres of Quaker Run Valley, New York

A) NUMBER of BREEDING PAIRS

to 3	1	670,656,196
.	9	
4	8	68,
6	7	23,23,
7	6	75,
8	5	06,
9	4	77,
14	3	89,67,24,11,10,
19	2	88,82,80,70,20,
29	1	88,81,79,61,60,58,52,38,11,09
31	9	10
32	8	8
33	7	9
34	6	0
37	5	760
(4)	4	6633
39	3	543220
33	2	886432
27	1	754200000
18	0	887665544443333211

(and an unknown number of zeroes)

37 + 4 + 39 = 80, hinges at 20h

Hinges: 10 and 184h

(Cont'd)

exhibit 24 of chapter 27 (Cont'd)

B) CALCULATIONS

Cut	log(0.5+cut) Cut	Diff	Count of pairs	Root of ratio	Converted midpoint	Gauss root Tabled	Matched
0.5	0.0						
		.30	2	2.58	-2.40	.52	3.14
1.5	.30						
		.30	5	4.08	-1.92	.65	3.94
3.5	.60						
		.30	9	5.48	-1.44	.78	4.72
7.5	.90						
		.30	9	5.48	-.96	.90	5.45
15.5	1.20						
		.30	8	5.17	-.48	.97	5.86
31.5	1.51						
		.30	13	6.58	.00	1.00	6.04
63.5	1.81						
		.30	6	4.48	.48	.97	5.86
127.5	2.11						
		.30	9	5.48	.96	.90	5.45
255.5	2.41						
		.30	11	6.05	1.44	.78	4.72
511.5	2.71						
		.30	4	3.65	1.92	.65	3.94
1023.5	3.01						
		.30	3	3.16	2.40	.52	3.14
2047.5	3.31						
(sum)				(52.17)		(8.64)	

Hinges

10	1.02			
184h	2.27	±2.88	.38	2.30*
midspread:	1.25	±3.36	.28	1.70**
2/midspread:	1.6	±3.84	.19	1.15***
(1.6)·(hinges):	1.63 and 3.64			
adjust:	2.64			

* corresponds to 1.6 pairs
** " " 0.9 "
*** " " 0.4 "

C) SOURCE

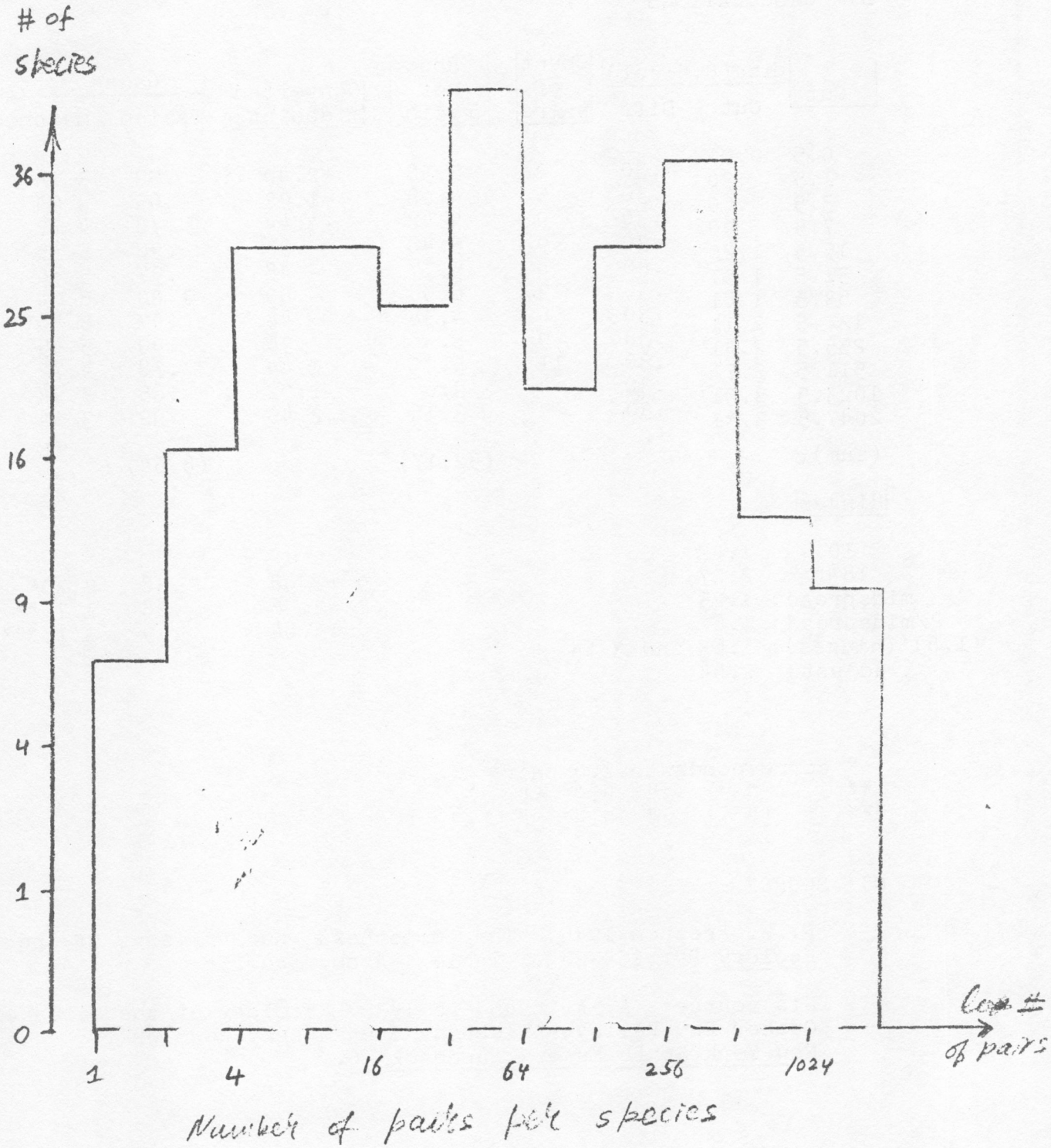
F. W. Preston 1945. The commonness, and variety, of species. Ecology 29: 254-283. Table 1-A on page 258.

His source: A. A. Saunders 1936. Ecology of the birds of Quaker Run Valley, Alleghany State Park, New York. New York State Museum Handbook 16.

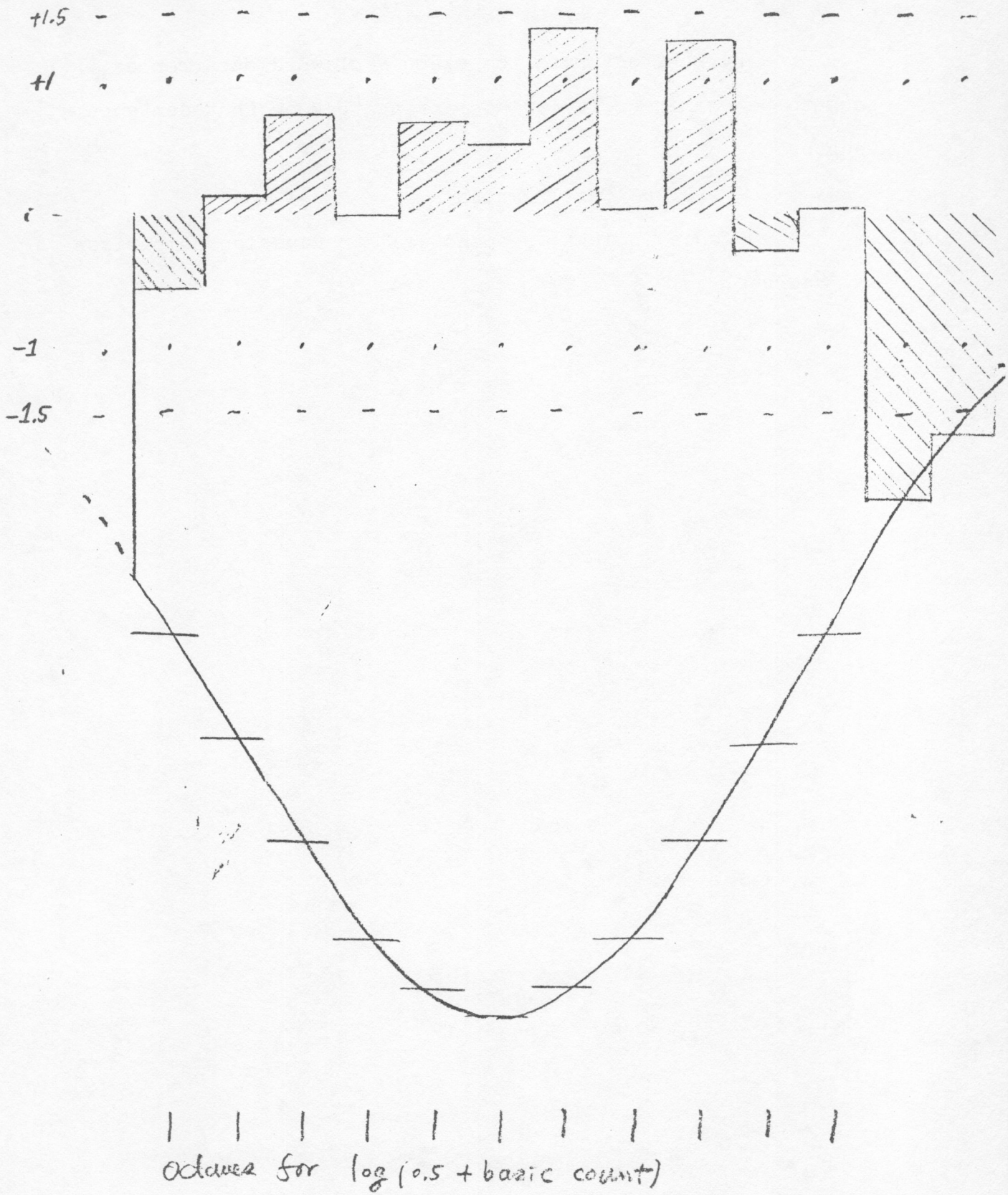
exhibit 25 of chapter 27 -- breeding birds

Breeding pairs of birds in Quaker Run Valley, N.Y.

(Rootogram by octaves, see exhibit 24)



Breeding Pairs of birds, Quaker Run Valley, N.Y.
Gauss-suspended rootogram for log (0.5 + basic count)
(see exhibit 24)



* sentence lengths *

By contrast with the example of word occurrences, soon to be examined, it is interesting to know that sentence lengths:

- cry out for logarithms.
- once in logs, respond well to Gaussian comparison curves.

Among statisticians at least, the study of distributions of sentence lengths was begun by Yule (1939, see also his book of 1944). In the same year, Williams pointed out the closeness of distributions of logs of sentence lengths to the Gaussian shape. (In this connection, Wake's paper (1957) on sentence-length distributions of Greek authors deserves attention.)

The point for us is not that logarithms of sentence lengths are distributed exactly in a Gaussian way, for they are not.

(As Wake says, p. 333: "If these distributions are transformed to a logarithmic scale we find that those derived from Aristotle have their upper tails missing whereas for Plato they approximate more nearly to a normal curve in the tail region.")

The point is, as Wake's words illustrate, that the actual distributions of the logarithms are close enough to Gaussian to make comparison with a Gaussian curve useful.

That this should be so for sentence length should not surprise us. Sentences of fewer than four or five words are very rare in connected prose, so that lengths of 4, 3, 2 and 1 will be successively less frequent, not more. Thus the distribution of sentence length is not J-shaped. At the other tail, sentences of much more than 100 words are quite rare but

not absent -- that tail is long -- when we look at raw counts -- but not extremely long. As an example of how logs help again, sentence lengths are very appropriate, but we can -- and will -- leave details for problems. (See section 27P.)

The frequencies of use of different words are of a quite different character. In a large corpus made up of connected prose, a very common word can occur many, many times. At the other tail, many words will appear only three times each, more twice each, still more once each. Indeed, of the 80 thousand words -- more or less -- to be found in an unabridged English dictionary, the vast majority will be absent. Thus there will be even more words with 0 occurrences than with one. At both tails the distribution of word frequencies is J-shaped. As a result, taking logs is not nearly so likely to bring these distributions under control.

27E. Reciprocal counts can be needed

In 1953, Josselson counted the occurrence of the Russian words in Pushkin's "The Captain's Daughter". Exhibit 27 gives the counts and examines them at log and reciprocal scales. Clearly we cannot look at these counts effectively in terms of logs. In terms of reciprocals, however, the situation is quite different.

The reciprocal examination made in panel B is an examination of

$$\frac{1}{0.5 + \text{occurrences}}$$

exhibit 27 of chapter 27 -- Pushkin's word

Examination of Josselson's Russian word counts

A) THE COUNTS

Basic counts appearing once each
1160,777,724
581,479,430
421,379,297
289,288,282
257,256,244
236,220,214
195,187,178
166,139,134
131,124,123
122,121,119
118,110,107
101, 98, 97
96, 95, 91
90

Stem and leaf	
8	99651
7	875332111
6	9833320
5	65553310
4	98887775
4	4333
4	111100000

Frequencies of individual counts			
39	5	19	7
38	6	18	6
37	4	17	20
36	4	16	17
35	7	15	12
34	1	14	22
33	5	13	20
32	9	12	33
31	8	11	34
30	10	10	39
29	5	9	41
28	6	8	79
27	3	7	82
26	6	6	114
25	5	5	146
24	3	4	238
23	7	3	433
22	7	2	847
21	10	1	2384
20	13	0	who knows?

B) EXAMINATION

log examination

Down to	There are	Diff
		1
1024	1	3
512	4	10
256	14	11
128	25	31
64	56	70
32	126	133
16	259	280
8	539	580
4	1119	1280
2	2399	2384
1	4783	

Reciprocal examination

Down to	There are	Diff	Div'd
		172	172
24	172	174	174
12	346	193	193
8	539	196	196
6	735	153	153
4.8	888	231	231
4	1119	431	215.5
3.4	"		
3	1512		
-	"	847	211.7
-	"		
-	"		
2	2399		
(skip 11)	"	2384	198.7
1	4783		

Notes: Entries in Down to column are 24/integer .
Entries in Div'd column are ratios of Diff to change in Down to .

(Cont'd)

exhibit 27 of chapter 27 (Cont'd)

C) ROOTOGRAM for 1/(0.5 + basic count) SUSPENDED from the RECTANGULAR

Count	Root	Fit	Residual	(Length)
172	13.11	13.88	-.77	(1)
174	13.19	13.88	-.69	(1)
193	13.89	13.88	.01	(1)
196	14.00	13.88	.12	(1)
153	12.36	13.88	-1.52	(1)
231	15.20	13.88	1.32	(1)
215.5	14.68	13.88	.80	(2)
211.7	14.55	13.88	.67	(4)
198.7	13.96	13.88	.08	(12)
(sum)	(124.94)			
(mean)	13.88			

D) PROBLEM

Repeat panel C for 2 or 3 other choices of a free start.

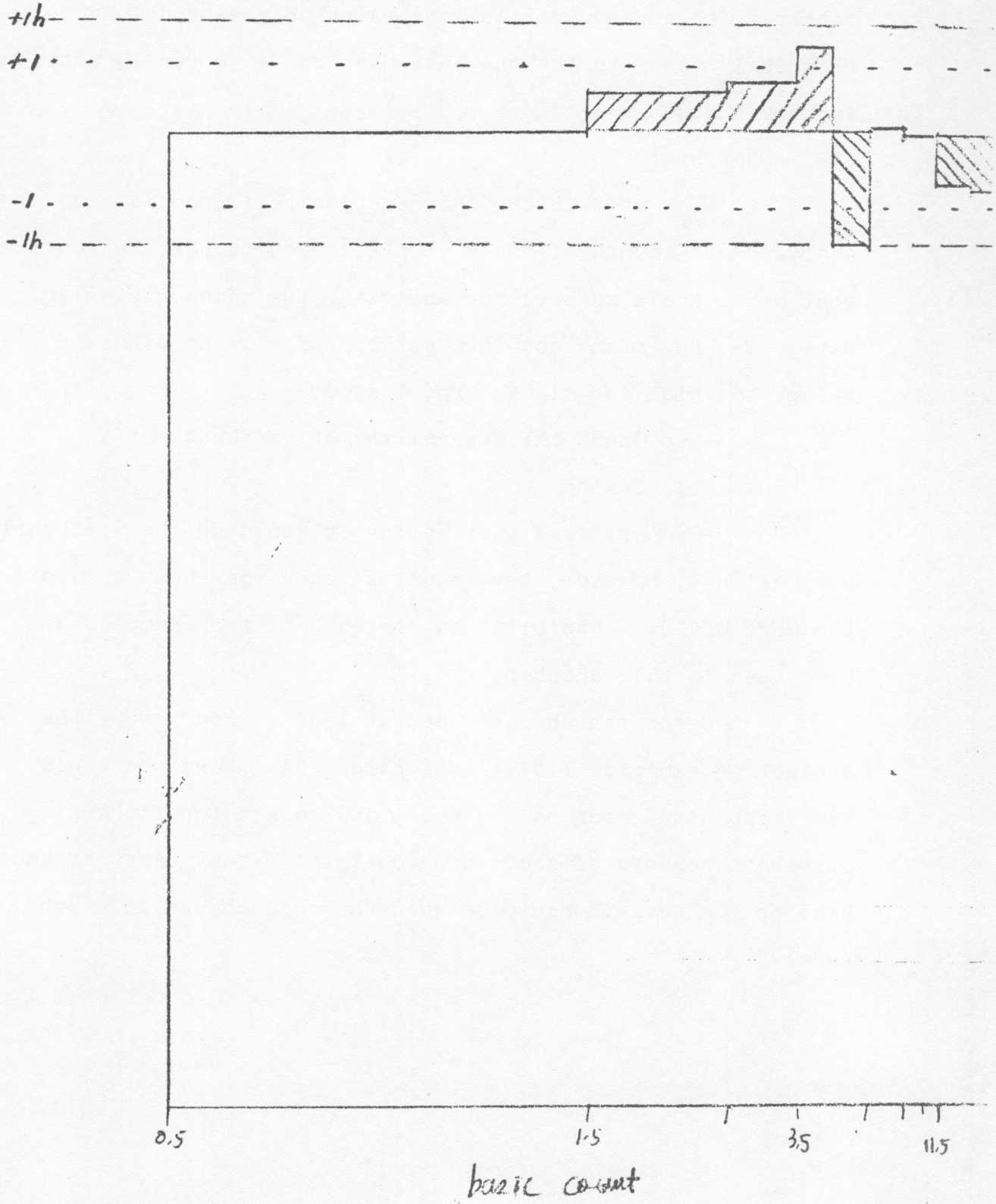
E) SOURCE

G. Herdan 1966. The advanced theory of language as choice and chance (Kommunikation und Kybernetik in einzel-darstellungen. Band 4). Springer, New York. Table 32 at page 97.

His source: H. Josselson 19--. The Russian Word Count.

exhibit 28 of chapter 27 -- Pushkin's words

Josselson's Russian word counts
Rootogram for started reciprocals suspended
from a rectangular comparison curve



which is surprisingly uniformly distributed. If we use a plot with equal area for each word, we are unlikely to do better. Indeed, under this constraint only expressions relatively close to the one just used can give really satisfactory plots.

27F. <u>What next?</u>

We stated in the introduction that once we have analyzed our counts of basic counts, we ought next to consider what other tools we have for analyzing the same or similar data. We know of no complete catalog of all the kinds of things to which one might turn, indeed:

- we doubt the possibility of existence of a complete catalog
- we are sure that having it would NOT be worthwhile.

We shall now, however, try to say a few words about some of the more obvious possibilities for each of the examples we have used in this chapter.

Seeds per pod per head of ladino clover were the subject of exhibits 1 to 4 in section 27A. Here one could have mapped the pods on the head, and looked for either different numbers of seeds per pod in different parts of the head or some relationship between the seed counts in nearby pods.

Scintillations per 1/8 minute period per experiment were the subject of exhibits 5 to 12 in section 27B. Here the 1/8 minute periods followed one another in stretches of some minutes duration. One could have looked at the results in time order, either to detect any relationship between counts at nearby times or to detect overall changes from one stretch to another.

Cells per square per haemacytometer were the subject of exhibits 13 and 14 in section 27B. Here the 400 counts apply to squares in 20x20 pattern. Student published, in his original paper, results of comparing rim squares with interior squares. One could also look either for relationships between nearby counts or for indications of patchiness.

Plants per quadrat per area of marsh studied are the subject of exhibits 15 to 19 in section 27C. If the quadrats were scattered -- hopefully randomly -- across a moderate area, the only obvious ways to do more would be to know the distance between each pair of quadrats. Given this, one can both look for either relationships between nearby counts or -- given a map -- for other indications of patchiness.

Mites per leaf per sample of leaves were the subject of exhibits 20 and 21 of section 27C. If we had mapped the leaf positions on the tree or trees involved - we would be ready to look at either reationships among counts on nearby leaves or systematic differences with height or orientation.

Words per sentence per piece of text was discussed as an example without exhibits. To the extent that complexity can be usefully and precisely defined, we could classify the sentences by their complexity -- or by their general purpose -- or by their grammatical character. (We leave other possibilities to the reader.)

Weeds per quadrat per area studied were the subject of exhibits 23 and 24 in section 27D. The same remarks apply as to the earlier quadrat example.

Breeding pairs per species per valley are the subject of exhibits 24 to 26 in section 27D. Here one might go further by breaking the valley down into parts, either blithely or in an attempt to separate habitats. Equally, one could break down species by feeding habits, by height preferences, or by order, family and genus.

Occurrences per word per text-sample are the subject of exhibits 27 and 28 in section 27E. Here one might want to separate the text into portions describing conversations, individuals, actions, rooms and buildings, landscapes, etc. Or one might want to separate words by grammatical class or linguistic origin. Or one might wish to group synonyms, actual or approximate.

27H. <u>How far have we come?</u>

Counting comes into its own when what we are analyzing consists of counts of occurrences of basic counts.

Examples show that we may want to express the basic counts:

- raw. (Very infrequently.)
- as double roots. (Most frequently.)
- as started logs. (Moderately frequently, usual start = 0.5.)
- as started reciprocals. (Perhaps frequently, usual start = 0.5.)

When we use double roots, we can find counts per unit base as

$$\frac{1}{2}\text{(doubled root of basic count)(count)}$$

except for basic counts of 0 or 1, where we use 2·count or 1.21·count , respectively.

Rootograms, often suspended, seem to be the displays of choice. The comparison curves for raw, root, and log expressions are usually Gaussian. For reciprocal expression a special display is usually preferred -- if we used a suspended rootogram we would use a rectangular comparison curve.

Techniques for very long-tailed distributions of counts appear in chapter 27x, following.

As always, the analysis of counts of basic counts is a step, and should encourage us to look further .

*****

27K. Sometimes other expressions are needed (background)

Sometimes things one would not think of as counts are in fact just that. In 1968, Chavez, Contreras G. and Hernandez D. reported on the duration of incubation of the ridley turtle on the coast of Tamaulipas. This might seem to be an amount, until we read that all the turtles emerged between 5:17 a.m. and 8:50 a.m. Clearly a count of days tells the main story. Exhibit 29 gives the counts. As this table suggests, even a rootogram is quite skew and one-long-tailed. Exhibit 30 shows this clearly.

Trying the usual forms of re-expression -- roots, logs, reciprocals -- makes very little difference. After all, 7/8 of the observations are between 50.7 and 56.3 days -- the larger is only 111% of the smaller.

If we are to do anything useful, we need to start close to 50 days. A handy place to start is 49 days. A handy re-expression is

$$\sqrt{-\ 49 + \text{time in days}}$$

Exhibit 31 carries out the calculations for an unequal-base rootogram. The plot is shown in exhibit 32 one summary of which was as follows:

- the number of turtles hatching after exactly 54 days is surprisingly large.
- the number of turtles hatching after exactly 52 days is surprisingly low.

exhibit 29 of chapter 27 -- turtle incubation

Duration of incubation of the ridley turtle

A) The COUNTS

Days	Count of turtles
50	77
51	122
52	10
53	321
54	725
55	180
56	162
57	21
58	14
59	6
60	9
61	14
(62)	
(63)	
64	1
65	1
(66)	
(67)	
(68)	
(69)	
70	1
(total)	(1664)

B) SOURCE

H. Chavez, M. Contreras G. and E. Hernandez D. 1968. On the coast of Tamaulipas. International Turtle and Tortoise Society Journal Vol. 2, No. 5 (Sept-Oct), pp. 16 to 19 and 27-34.

exhibit 30 of chapter 27 -- turtle incubation

Duration of incubation for the ridley turtle
at Tamaulipas Bay (see exhibit 29)
(raw days interval, compare exhibit 32)

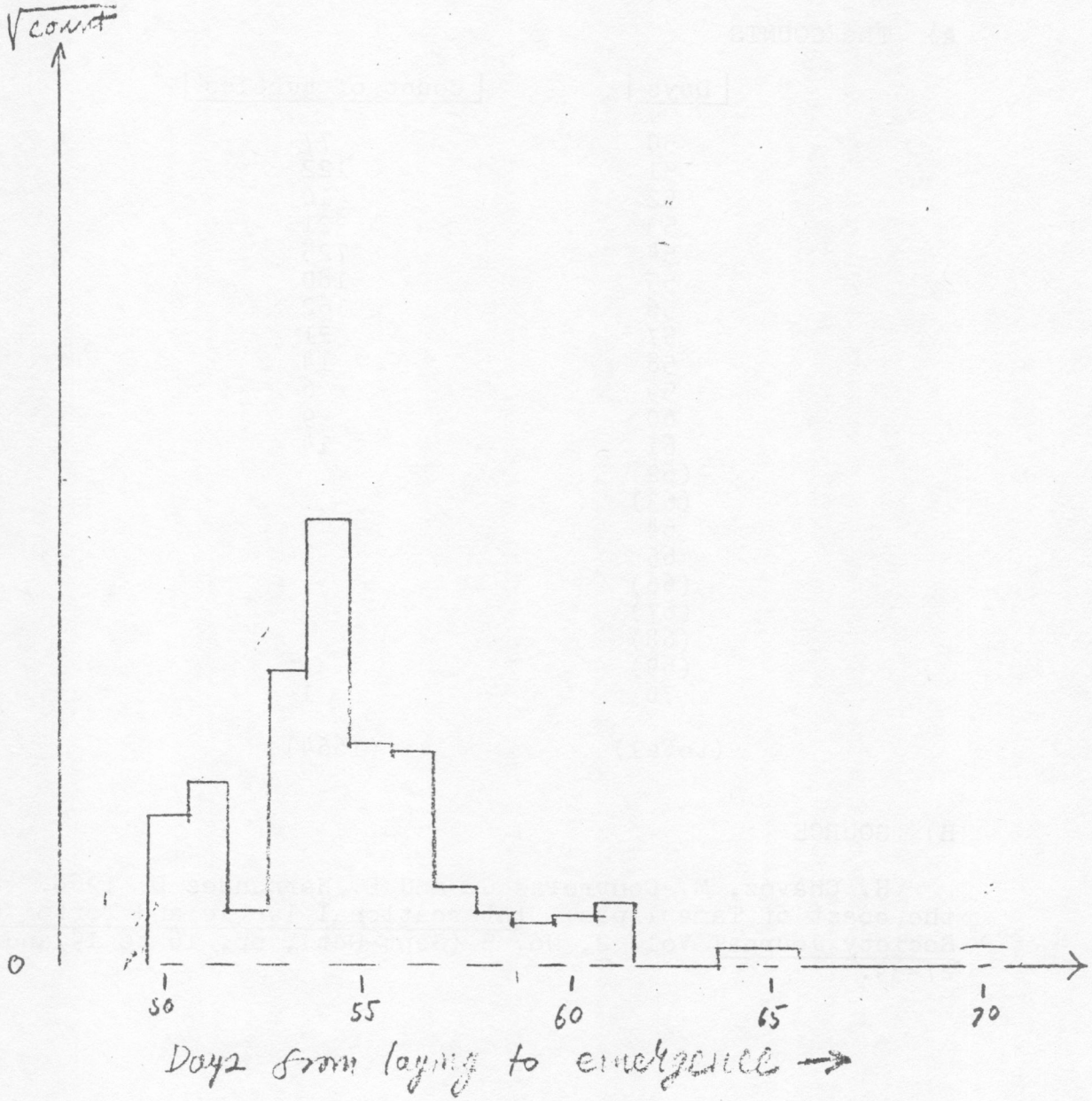


exhibit 31 of chapter 27 -- turtle incubation

$\sqrt{-\ 49 \text{ days} + \text{incubation period}}$ for the ridley turtle

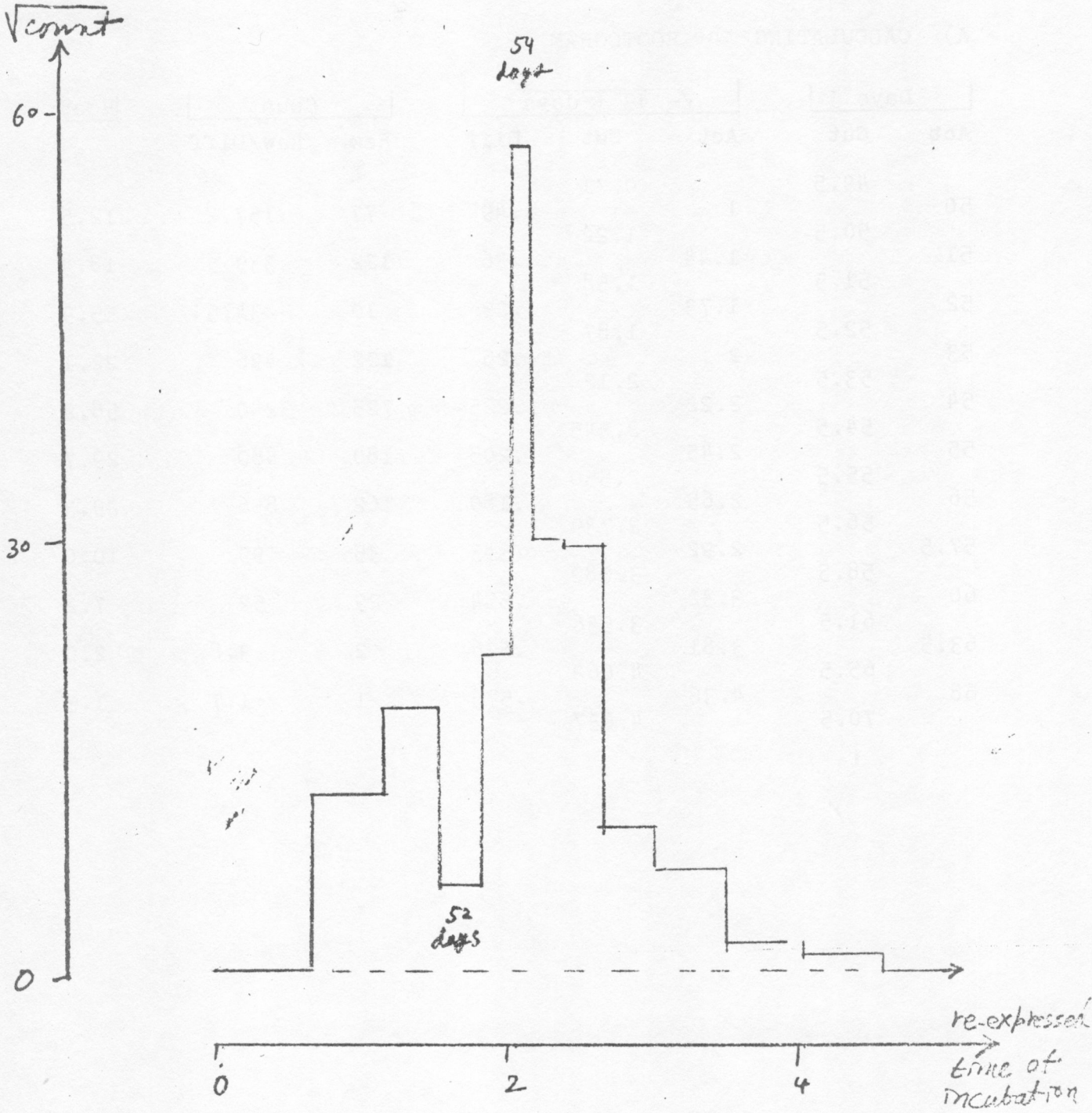
(see exhibit 29)

A) CALCULATING the ROOTOGRAM

Days		$\sqrt{-\ 49 + \text{days}}$			Count		Root
Act	**Cut**	**Act**	**Cut**	**Diff**	**Raw**	**Raw/Diff**	
	49.5		0.71				
50		1		.49	77	157.2	12.5
	50.5		1.22				
51		1.44		.36	122	339	18.4
	51.5		1.58				
52		1.73		.29	10	34.5	5.9
	52.5		1.87				
53		2		.25	122	496	22.1
	53.5		2.12				
54		2.22		.225	725	3240	56.9
	54.5		2.345				
55		2.45		.205	180	880	29.7
	55.5		2.550				
56		2.65		.189	162	856	29.3
	56.5		2.739				
57.5		2.92		.343	35	99	10.0
	58.5		3.082				
60		3.32		.554	29	52	7.2
	61.5		3.536				
63.5		3.81		.526	2	3.8	2.0
	65.5		4.062				
68		4.36		.575	1	1.7	1.3
	70.5		4.637				

exhibit 32 of chapter 27 -- turtle incubation

Duration of incubation for the ridley turtle at Tamaulipas Bay
(re-expressed, compare exhibit 30)



- the number of turtles hatching after exactly 49 days may be low.

- otherwise, the distribution of $\sqrt{-\ 49 + \text{incubation period}}$ seems quite usual.

Since the authors state that there were days when some turtles hatched before the observers got there, and that no attempt was made to include such turtles in the count, we may suspect that this is the explanation of the deficit at 52 days -- and, possibly, of the possible deficit at 49 days.

The moral of this example is NOT that

$$\sqrt{-\ \text{constant} + \text{count}}$$

is always worth trying.

The proper moral is simple:

- if re-expression is needed, and the standard re-expressions don't work, try something that does.

27L. Poisson distributions

The usual ideal basis of comparison for counts of basic counts is a family of standard patterns called Poisson distributions. For any specified arithmetic mean count, there is exactly one Poisson distribution. Anyone can do a more or less careful job of fitting a Poisson distribution to any set of counts of basic counts. The fit may be good or bad. Once this is done, we can look at the corresponding rootogram -- or rootobar plot -- and see how good or bad a fit we have.

For our purposes, we will almost always do well enough by using double roots for the basic counts and a Gaussian comparison curve. The proof of this appears when we see how close to matching the Gaussian comparison even those Poisson distributions we would think unlikely hopes come when we make suspended rootograms using double roots of the basic counts.

Exhibit 33 shows the arithmetic for the Poisson distribution of mean basic count 2. If we interpolate hinges, and use the conversion that goes with the interpolated hinges, we find a largest residual -- in $\sqrt{\text{fraction}}$ -- of about .04 -- see panel A. Exhibit 34 shows the corresponding suspended rootogram. This pattern of residuals is one we can reduce by changing the conversion. A little cut-and-try brings us to the comparison curve of panel C of exhibit 33. The largest residual is now only ±.01.

Exhibit 35 shows the improved suspended rootogram.

Suppose we had a sample of 10,000 counts distributed in as exact accord with our Poisson distribution as only 10,000 values can be. There would then be 1353 zero counts. The square root of 1353 is 36.8, and if we allow ±1 we get 35.8 and 37.8. Dividing by 1.414 = $\sqrt{2}$, the corresponding values of $\sqrt{\text{ratio}}$ are 26.0, 25.3, and 26.7. The fit of panel C would give a fitted value of 25.0, which falls at -1.4 in $\sqrt{\text{count}}$. Thus we

exhibit 33

Fitting a comparison curve to the Poisson distribution with mean 2
-- Poisson fraction

A) FITTING ARITHMETIC

Double roots			Poisson fraction	√ of ratio*	Conversion	Relative height		
(Center)	Cut	Diff				Tabled	Fitted	Diff
	0.0000							
1.41		2.0000	.1353	.260	-2.61	.46	.29	-.03
	2.0000							
2.45		.8284	.2707	.571	-.96	.90	.56h	.00h
	2.8284							
3.17		.6357	.2707	.653	.17	1.00	.62h	.03
	3.4641							
3.74		.5359	.1804	.580	1.12	.86	.54	.04
	4.0000							
4.21		.4721	.0902	.437	1.92	.65	.41	.02
	4.4721							
4.61		.4268	.0361	.291	2.62	.45	.28	.01
	4.8990							
5.10		.3925	.0120	.175	3.28	.30	.19	-.01h
	5.2915							
5.47		.3654	.00343	.098	3.88	.18	.11h	-.01h
	5.6568							
5.83		.3432	.000859	.050	4.43	.11	.07	-.02
	6.0000							
6.16		.32460	.000191	.024	4.88	.07	.04h	-.02
	6.3246							
6.48		.3087	.000038	.010	5.51	.03	.02	-.01
	6.6332							
6.78		.2950	.000007	.005	6.00	.01	.00h	.00
	6.9282							
7.07		.2829	.000001	.002				
	7.2111							
		(sum)		(3.156)		(5.025)	(3.15h√)	

* √ of ratio of Poisson fraction to difference of double roots.

** see panel B < slide-rule accuracy).

(Cont'd)

exhibit 33 (Cont'd)

B) INTERPOLATION for HINGES and CONVERSION

Double-root cut	Cumulated fraction	Interpolate	Trial of tilt
2.000	.1353		
	(.2500)	2.352	3.56
2.828	.4060		
*	*		
*	*		
4.000	.1428		
	(.2500)	3.671	5.56
3.464	.3232		
	(Diff)	(1.319)	
	2/Diff	1.51	

Conversion: 1.51(double root) - 4.56

C) A BETTER FIT -- found by cut and try

			Relative height		
Count	√ratio	Conversion*	Tabled	Fitted	Diff
0	.260	-2.91	.38	.25	+.01
1	.571	-1.14	.86	.56	+.01
2	.653	.09	1.00	.65h	-.00
3	.580	1.06	.88	.57h	+.01
4	.437	1.90	.66	.43	+.00h
5	.291	2.70	.43h	.28h	+.00h
6	.175	3.37	.27h	.18	-.00h
7	.098	4.00	.16	.10h	-.00h
8	.050	4.61	.09	.06	-.01
9	.024	5.17	.05	.03h	-.01
10	.010	5.74	.02	.01h	-.00h
11	.005	6.25	.01	.01	-.01
12	.002	6.72	.00h	.00h	-.00h
			(4.82h)		

Conversion used: 1.7(double root) + 4.9.

D) SOURCE -- for Poisson fractions

E. S. Pearson and H. O. Hartley 1966 (1954). Biometrika Tables for Statisticians, volume 1. Table 39 at page 218 of 3rd edition.

exhibit 34 of chapter 27 -- Poisson fractions

First try at fitting Poisson Distribution (when mean is 2)

(Gaussian comparison, see exhibit 33)

(compare exhibit 35)

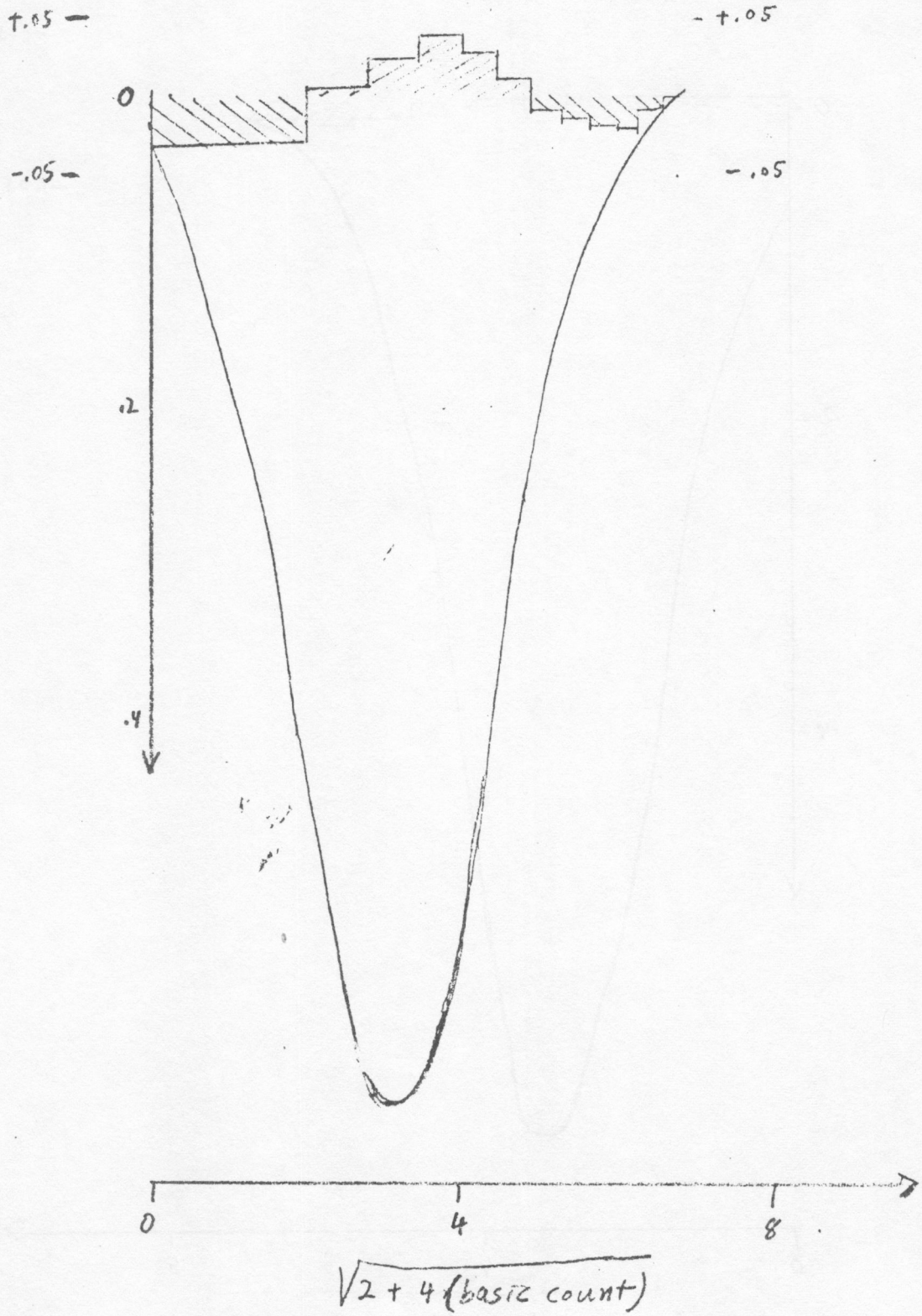
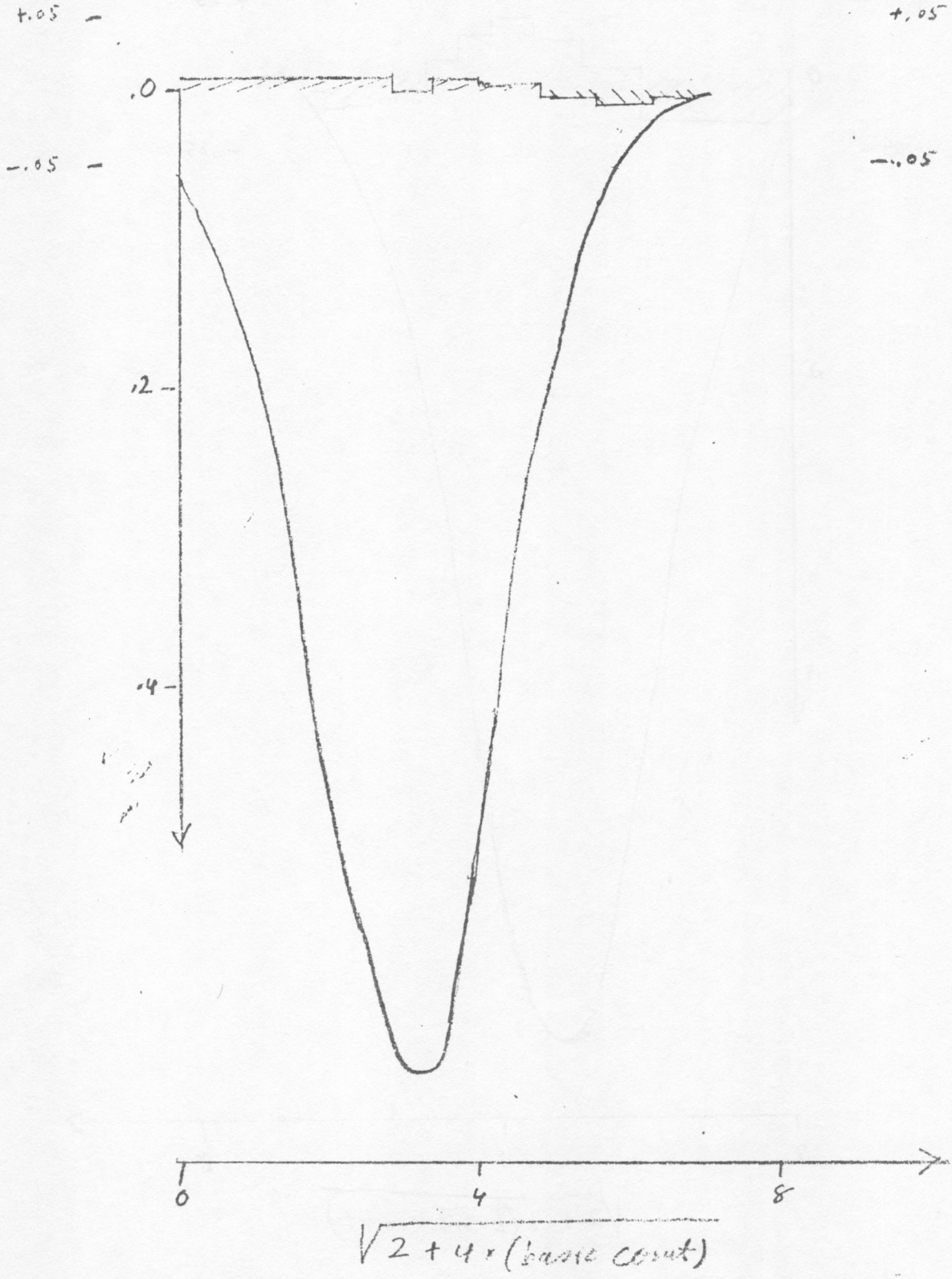


exhibit 35 of chapter 27 -- Poisson fractions

Better fit to Poisson distribution (when mean is 2)
(Gaussian comparison, see exhibit 33)
(compare exhibit 34)



would judge this residual -- and in fact all the residuals for a panel C fit -- as quite reasonable since -1.4 is between -1.5 and +1.5.

Unless we have more than 10,000 counts we do not have to worry about the approximate character of Gauss-suspended rootograms using doubled-roots for Poisson counts.

So long as our batch of basic counts does not exceed 1000 -- even for so small a mean basic count as 2 -- the Gaussian comparison curve will be a very adequate approximateion to a Poisson comparison. With larger counts, it may still be quite close enough. True, we will, once in a while, find it useful to actually fit a Poisson distribution, but maybe this will happen only once in a long, long time.

*****

27P. Problems for chapter 27

1. Vansell (reference in exhibit 3) gave seed counts for three other open pollinated heads of Ladino clover, namely:

Seeds:	0	1	2	3	4	5	6
pods:	8	10	12	27	33	13	1
pods:	8	15	20	28	15	5	0
pods:	18	28	41	39	15	2	0

Find and plot suspended rootograms for at least one of these. Compare with exhibit 4. Comment.

2. In 1944, Flacker and Brischle reported the number of *Ribes* plants (currants and gooseberries, dangerous to white pine as alternating hosts for the white pine blister rust) in 64 2.5-acre plots in Clearwater National Forest. Exhibit 36 gives the counts. Prepare a rootogram, using double roots for the basic counts. Compare with exhibit 19. Comment.

3. In 1946, Cole published counts of fleas (*Xenopsylla Cheopis*) on 209 live rats. Exhibit 37 gives the data. Plot an appropriate rootogram. Comment.

4. In the same paper, Cole published counts of isopods (*Trachelipus rathkei*) found under each of 122 boards. Exhibit 38 gives the data. Plot an appropriate rootogram. Comment.

exhibit 36 of chapter 27

Counts of Ribes in 2.5 Acre Block
in Clearwater National Forest

A) DATA

```
Z | 0000
0 | 33211234123
0 | 596598
1 | 0721159
2 | 888738
3 | 4272
4 | 01810
5 | 6468655
6 | 77
7 | 27
8 | 97
9 |
1 | [83,53,43,93,
2 | [57,67,45,67,
```

B) SOURCE

S. B. Flacker and H. A. Brischle 1944. Measuring the local distribution of Ribes. Ecology 28, 288-303. Figure 3 on page 290.

exhibit 37 of chapter 27

Counts of Rat Fleas on Live Rats in Mobile, Alabama

A) THE DATA -- in two sections, first counts and then stem-and-leaf for basic counts

Fleas per rat	Count
0	63
1	32
2	30
3	18
4	8
5	8
6	11
7	3
8	5
9	5
.	
.	
.	

stem-and-leaf

1	000011222344567889
2	1378
3	
4	11
5	
6	11

B) SOURCE

L. C. Cole 1946. A theory for analyzing contagiously distributed populations. Ecology 27: 329-341. Table VI on page 338.

exhibit 38 of chapter 27

Counts of An Isopod (Trachelipus Rathkei)
Found Under Each of 122 Boards

A) THE DATA -- in two sections, first counts and then stem-and-leaf for basic counts

Isopods	Count
0	28
1	28
2	14
3	11
4	8
5	11
6	2
7	3
8	3
9	3

stem-and-leaf

1|00011344577

B) SOURCE

L. C. Cole 1946. A theory for analyzing contagiously distributed populations. Ecology 27:329-341. Table V at page 337.

His source: L. C. Cole Ecological Monographs 16.

5. In 1951 Morgan, MacLeod, Anderson and Bliss published counts of the number of bacterial clumps in each of 399 microscope fields in a milk film. Exhibit 39 gives the data and a negative binomial fit given by Bliss in 1953. Calculate and plot the two rootograms analogous to those in exhibit 21. Comment.

6. In 1953, Bliss published three different fits to counts of lice of all stages on the heads of Hindu male prisoners in Cannamore, South India. Exhibit 40 gives the observations and the fits. Calculate and plot the three suspended rootograms using raw basic counts. What can you say about their comparison? Do the same for the suspended rootograms using double roots of basic counts. What now?

7. In 1953, Beall and Rescia fitted three theoretical distributions to Archibald's data on the number of plants of Salicornia stricta. Exhibit 41 gives the numbers. Draw the suspended rootograms. Which fit seems better? How much better?

exhibit 39 of chapter 27

Counts of Bacterial Clumps for Microscope Fields in a Milk Film, and a Fit

A) THE DATA

Clumps per field	Count of fields	Fitted count
0	56	64.2
1	104	90.3
2	80	82.7
3	62	62.1
4	42	41.6
5	27	25.8
6	9	15.1
7	9	8.5
8	5	4.7
9	3	2.5
10	2	1.3
≥11	0	1.2

SOURCE

M. E. Morgan, P. MacLeod, E. O. Anderson, and C. I. Bliss 1951. A sequential procedure for grading milk by microscopic counts. Storrs Agricultural Experiment Station Bulletin 276.

Used by: C. I. Bliss 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176-190. Table 2 on page 186.

exhibit 40 of chapter 27

Buxton's Data on Lice of All Stages on the Heads of Hindu Male Prisoners in Cannamore, South India (1937-1939) and Three Fits

Lice per head	Count of heads			
	Obs'd	Fit A	Fit A'	Fit B
0	612	612.0	612.0	XXX
1	106	86.5	90.5	107.2
2	50	48.5	50.8	57.8
3	29	33.9	35.5	34.7
4	33	26.1	27.3	25.6
5	20	21.2	22.1	22.2
6	14	17.8	18.5	16.6
7	12	15.3	15.8	14.0
8	18	13.4	13.8	12.1
9	11	11.9	12.2	10.6
10	11	10.6	10.9	9.4
11	3	9.6		8.4
12	10	8.7		7.6
13	8	8.0		6.9
14	6	7.4		6.3
15	3	6.8		5.8
16	6	6.3		5.4
17	7	5.9		5.0
18	4	5.5		4.6
19	7	5.1		4.3
20	7	4.8		4.1
21	3	4.5		3.8
22-23	8	8.3		7.0
24-25	7	2.4		6.2
26-27	10	6.6		5.6
28-29	0	6.0		5.0
30-32	2	7.9		6.7
33-35	7	6.9		5.9
36-38	9	6.0		5.2
39-41	5	5.3		4.6
42-45	5	6.1		5.3
46-49	7	5.2		4.6
50-	27	37.5		37.5

SOURCE

C. I. Bliss 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176-190. Table 8 on page 194.

His source: F. J. Anscombe 1950. Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37:358-382.

Anscombe's sources: C. B. Williams 1944. J. Ecology 32:1-44.
P. A. Buxton 1940. Parasitology 32:296-

exhibit 41 of chapter 27

Three Fittings by Beall and Rescia

	Fitted counts					Fitted roots*		
Count	n=0	n=5	n→∞		Root*	n=0	n=5	n→∞
0	4	10.4	6.9	6.0				
1	3	4.0	6.7	6.9				
2	8	6.6	7.7	8.0				
3	13	7.9	8.2	8.5				
4	11	8.2	8.3	8.6				
5	9	8.0	8.1	8.3				
6	8	7.6	7.7	7.8				
7	10	7.1	7.0	7.1				
8	3	6.4	6.3	6.3				
9	3	5.7	5.5	5.5				
10	8	5.0	4.8	4.7				
11	3	4.2	4.1	4.0				
12	4	3.5	3.4	3.3				
13	4	2.9	2.8	2.7				
14	0	2.4	2.3	2.2				
15	3	1.9	1.9	1.8				
16	0	1.5	1.5	1.4				
17	0	1.2	1.2	1.1				
18	1	.9	1.0	.9				
≥19	3	2.6	2.6	2.9				

* Use doubled roots.

Source. G. Beall and R. R. Rescia 1953. A generalization of Neyman's contagious distributions. Biometrics 9:354-386. Table VI at page 373.

Data: E. E. A. Archibald 1948. Plant populations I. A new application of Neyman's contagious distribution. Annals of Botany 12:221-235.

8. Do the problem of panel D of exhibit 27.

9. Turn to exhibit 47 of chapter 25 (for problem 4 of that chapter) and make suspended rootograms for log (sentence length) for both sets of data. Comment.

(Optional addition). Make a suspended rootogram for log (0.5 + sentence length) for one of these sets. Compare and contrast the two suspended rootograms for the same set of data.

10. Exhibit 42 gives the fractions for Poisson distributions with means 1, 4 and 9. Choose one, convert to double roots and fit a comparison curve with care. Draw the suspended rootogram, and calculate about how large a batch of basic counts are needed to make the deviation between Poisson distribution and comparison curve as much as ±0.3.

11. Do the problem of panel B of exhibit 15.

exhibit 42 of chapter 27

Three More Poisson Distributions

A) FRACTION

Count	Poisson fraction		
	λ=1	λ=4	λ=9
0	.3679	.0183	.000123
1	.3679	.0733	.001111
2	.1839	.1465	.004998
3	.0613	.1954	.01499
4	.0133	.1954	.03374
5	.00307	.1563	.06073
6	.00051	.1042	.09109
7	.000073	.05954	.1171
8	.000009	.02977	.1318
9		.01323	.1318
10		.00529	.1186
11		.00192	.0972
12		.000642	.0728
13		.000197	.0504
14		.000056	.0324
15		.000015	.0194
16		.000004	.0109
17		.000001	.0058
18			.00289
19			.00137
20			.000617
21			.000264
22			.000108
23			.000042
24			.000016
25			.000006

B) SOURCE

E. S. Pearson and H. O. Hartley 1966 (1954). Biometrika Tables for Statisticians, Volume 1. Table 39 on pages 218, 219, 221 of 3rd edition.